Provenance-Enabled Data Exploration and Visualization Tutorial

Erik Anderson
Juliana Freire
David Koop
Emanuele Santos
Claudio Silva

Presentation VIS

“Study of a Numerically Modeled Severe Storm”, NCSA, UIUC

Presentation VIS

“Study of a Numerically Modeled Severe Storm”, NCSA, UIUC

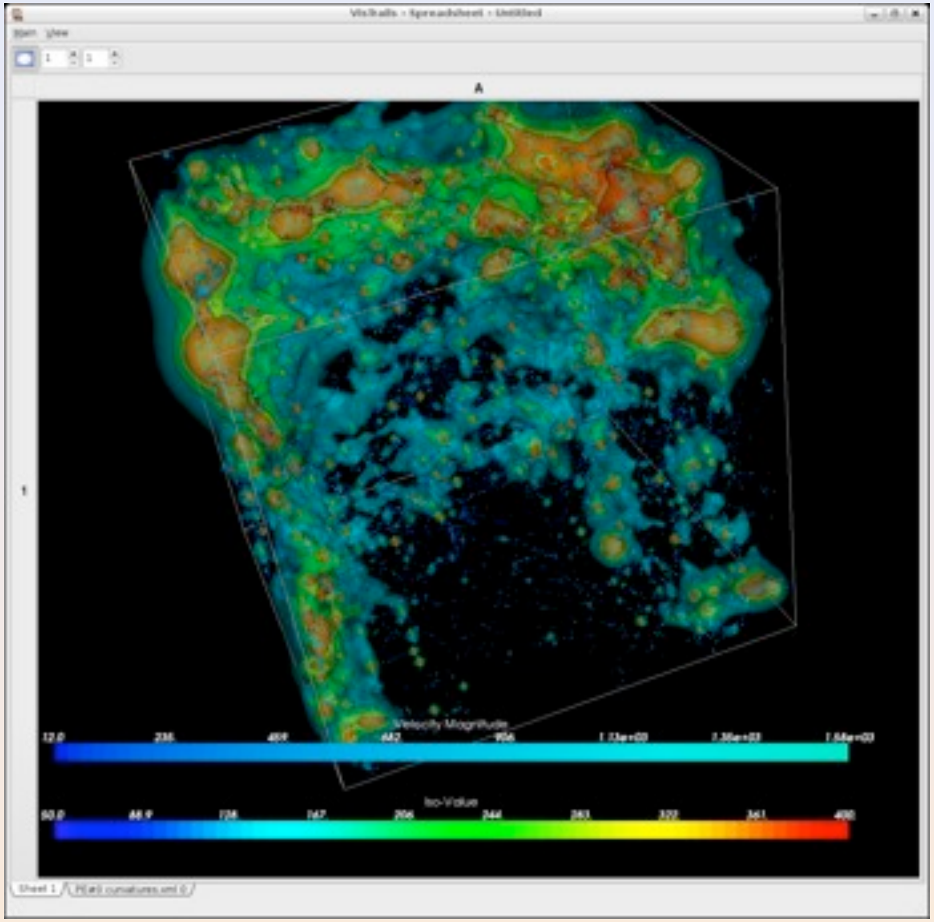
Presentation VIS

“Fusion Simulation Visualization” Kruger, Sanderson, et al

Presentation VIS

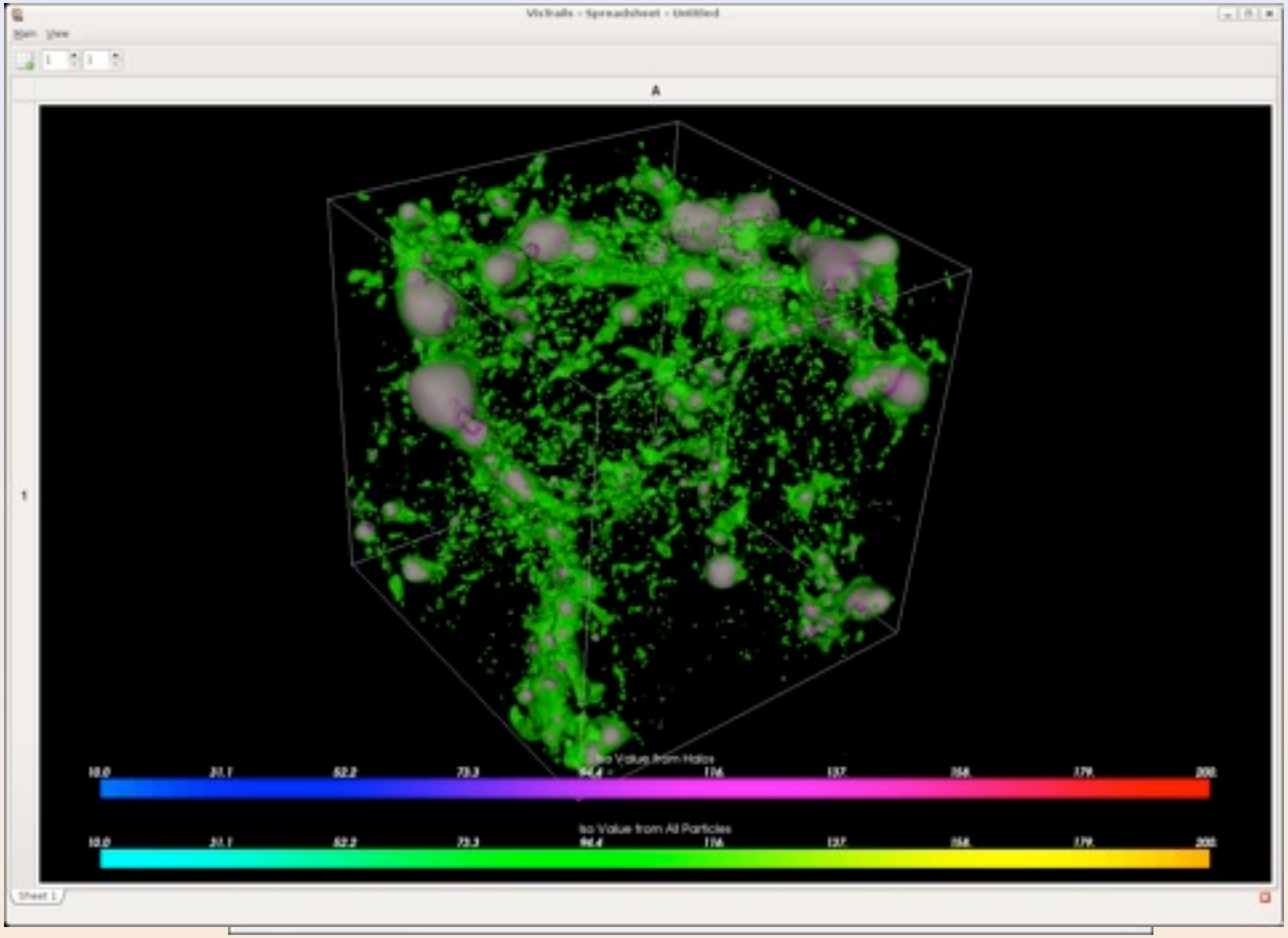
“Fusion Simulation Visualization” Kruger, Sanderson, et al

Exploratory VIS



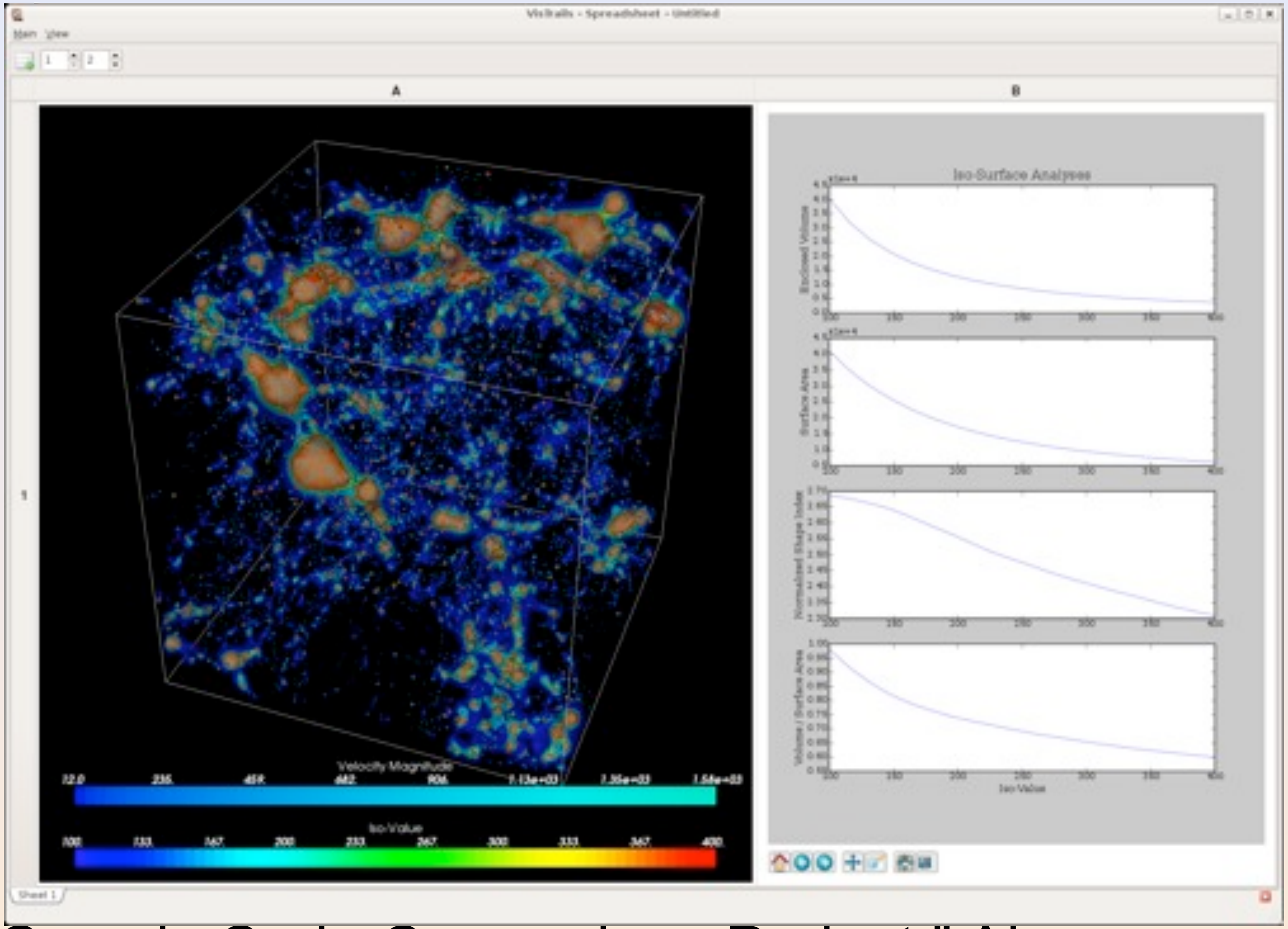
“The Cosmic Code Comparison Project,” Ahrens, Anderson, Heitmann, Habib, et al

Exploratory VIS



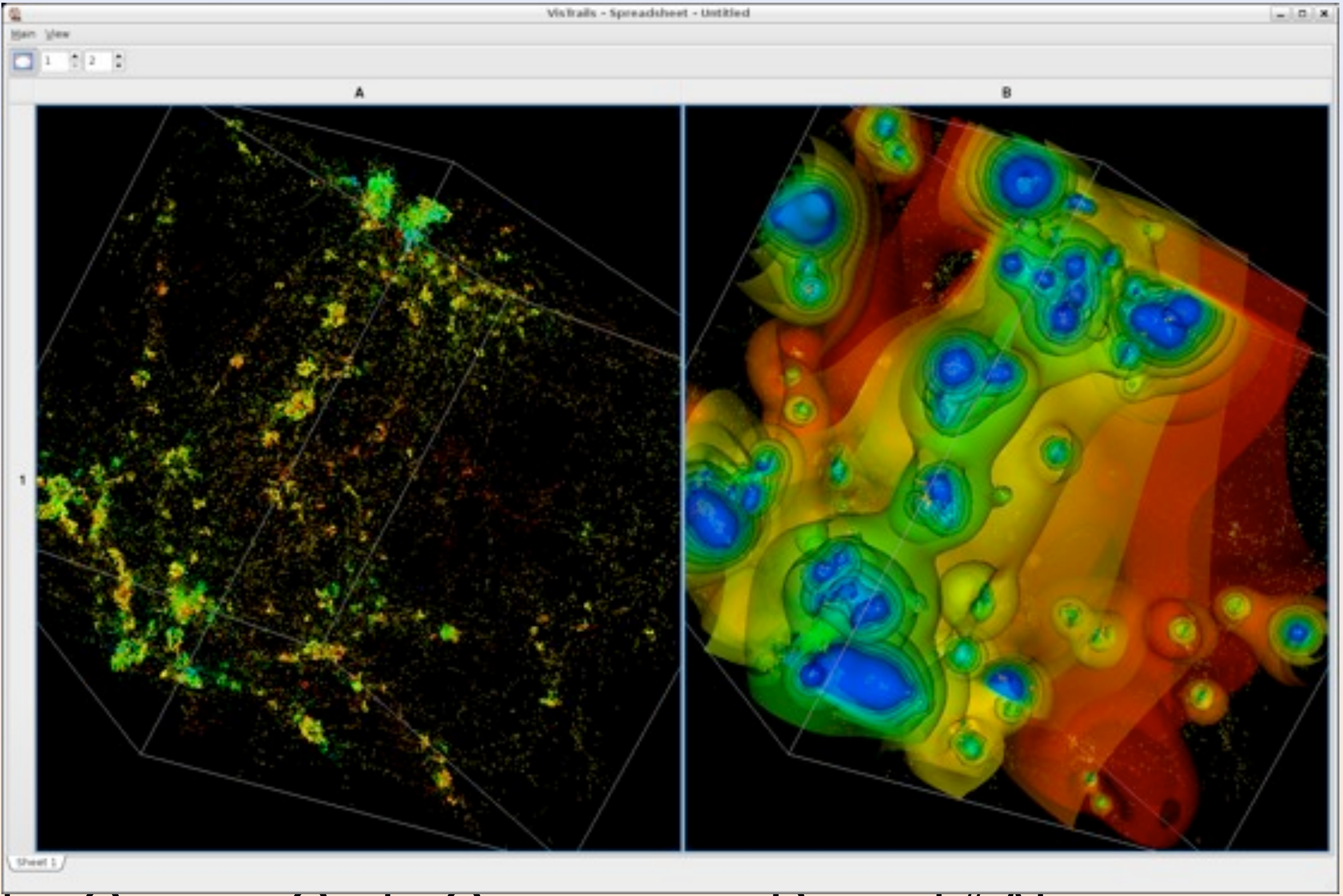
“The Cosmic Code Comparison Project,” Ahrens, Anderson, Heitmann, Habib, et al

Exploratory VIS



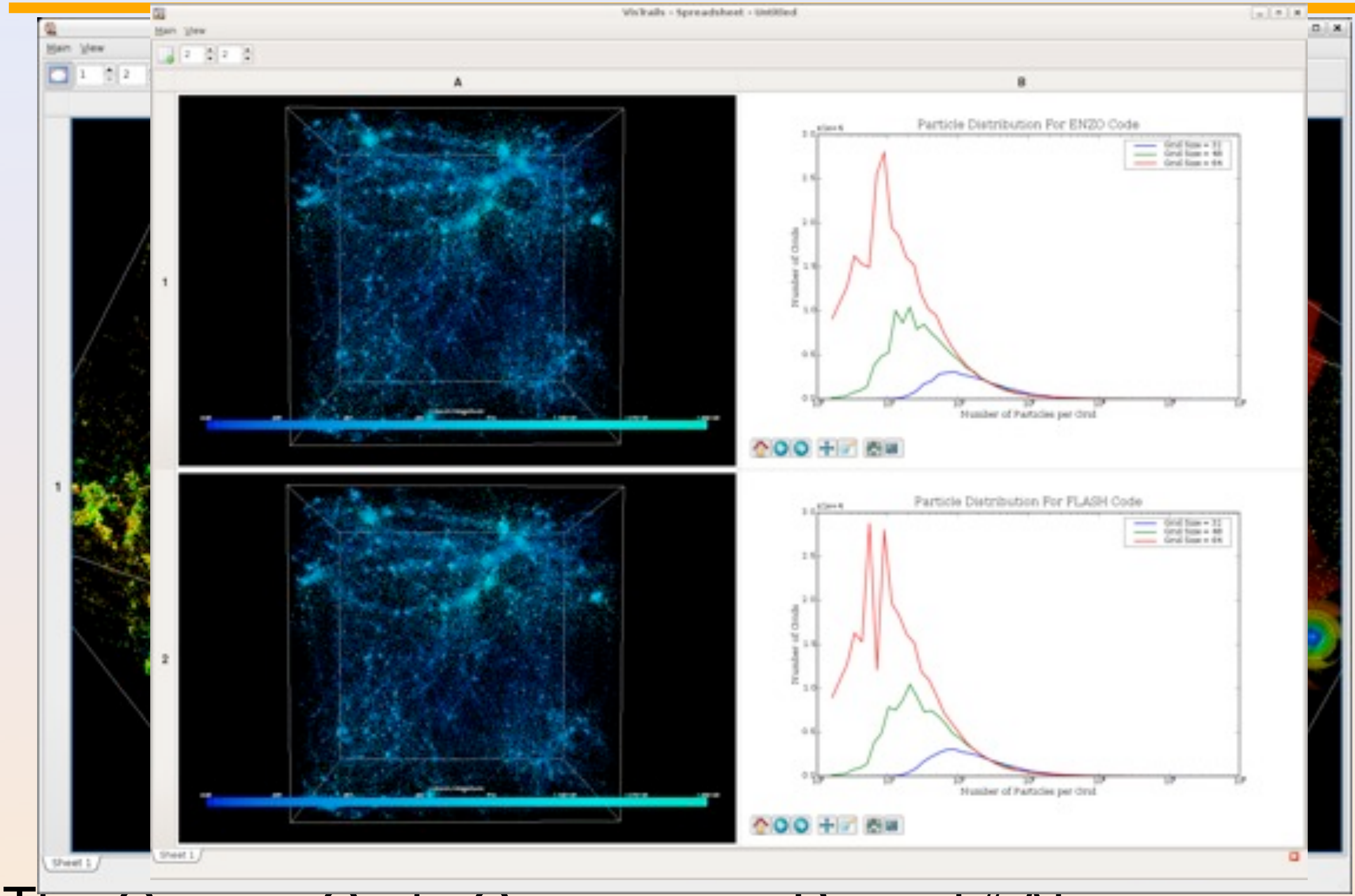
“The Cosmic Code Comparison Project,” Ahrens, Anderson, Heitmann, Habib, et al

Exploratory VIS



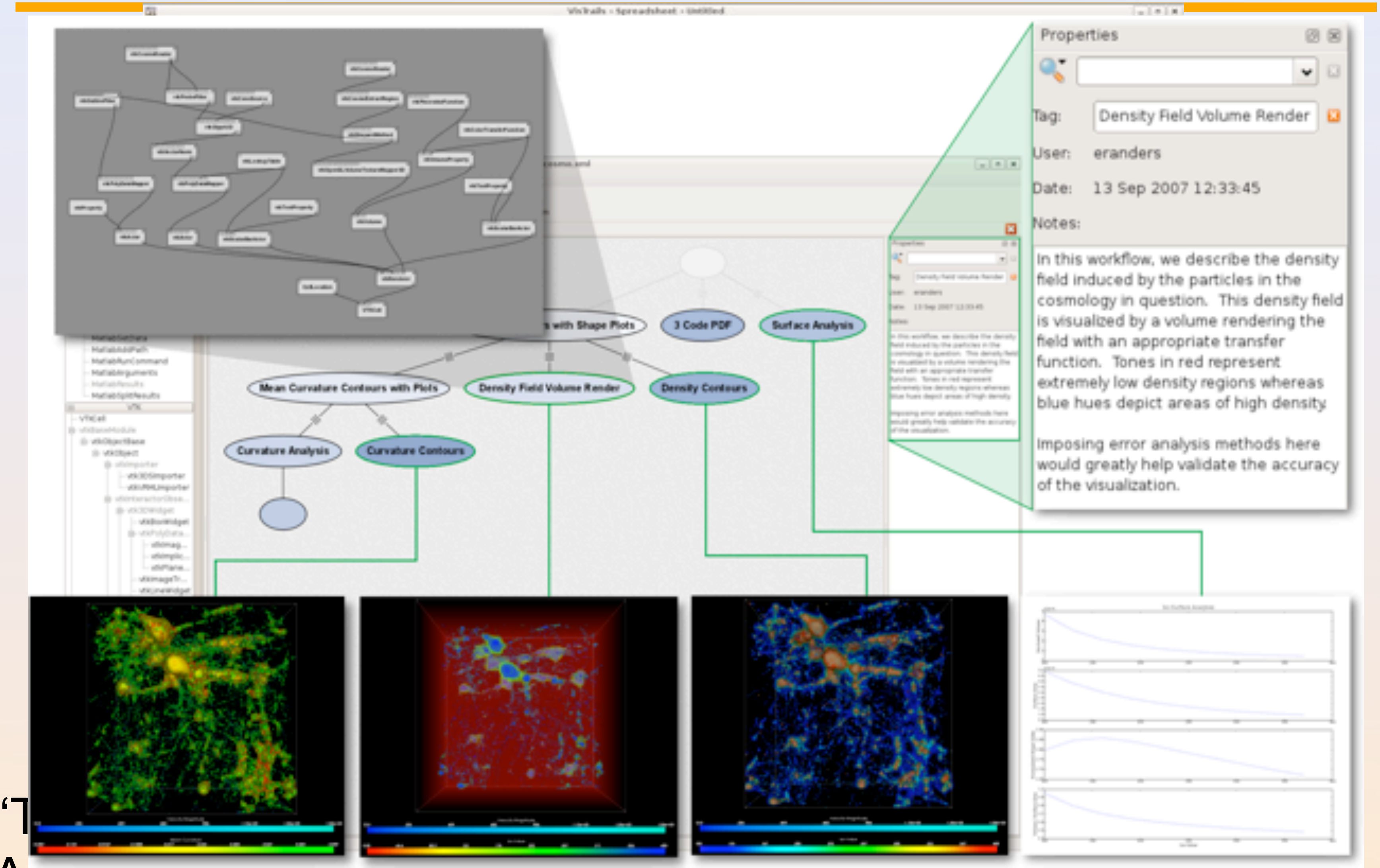
"The Cosmic Code Comparison Project," Ahrens, Anderson, Heitmann, Habib, et al

Exploratory VIS


VisTrails - Spreadsheet - Untitled
Main View
A
B
1
2
Particle Distribution For ENZO Code
Grid Size = 32
Grid Size = 48
Grid Size = 64
Number of Grids
Number of Particles per Grid
Particle Distribution For FLASH Code
Grid Size = 32
Grid Size = 48
Grid Size = 64
Number of Grids
Number of Particles per Grid
Sheet 1
Sheet 1

"The Cosmic Code Comparison Project," Ahrens, Anderson, Heitmann, Habib, et al

Exploratory VIS



Tag: Density Field Volume Render

User: eranders

Date: 13 Sep 2007 12:33:45

Notes:

In this workflow, we describe the density field induced by the particles in the cosmology in question. This density field is visualized by a volume rendering the field with an appropriate transfer function. Tones in red represent extremely low density regions whereas blue hues depict areas of high density.

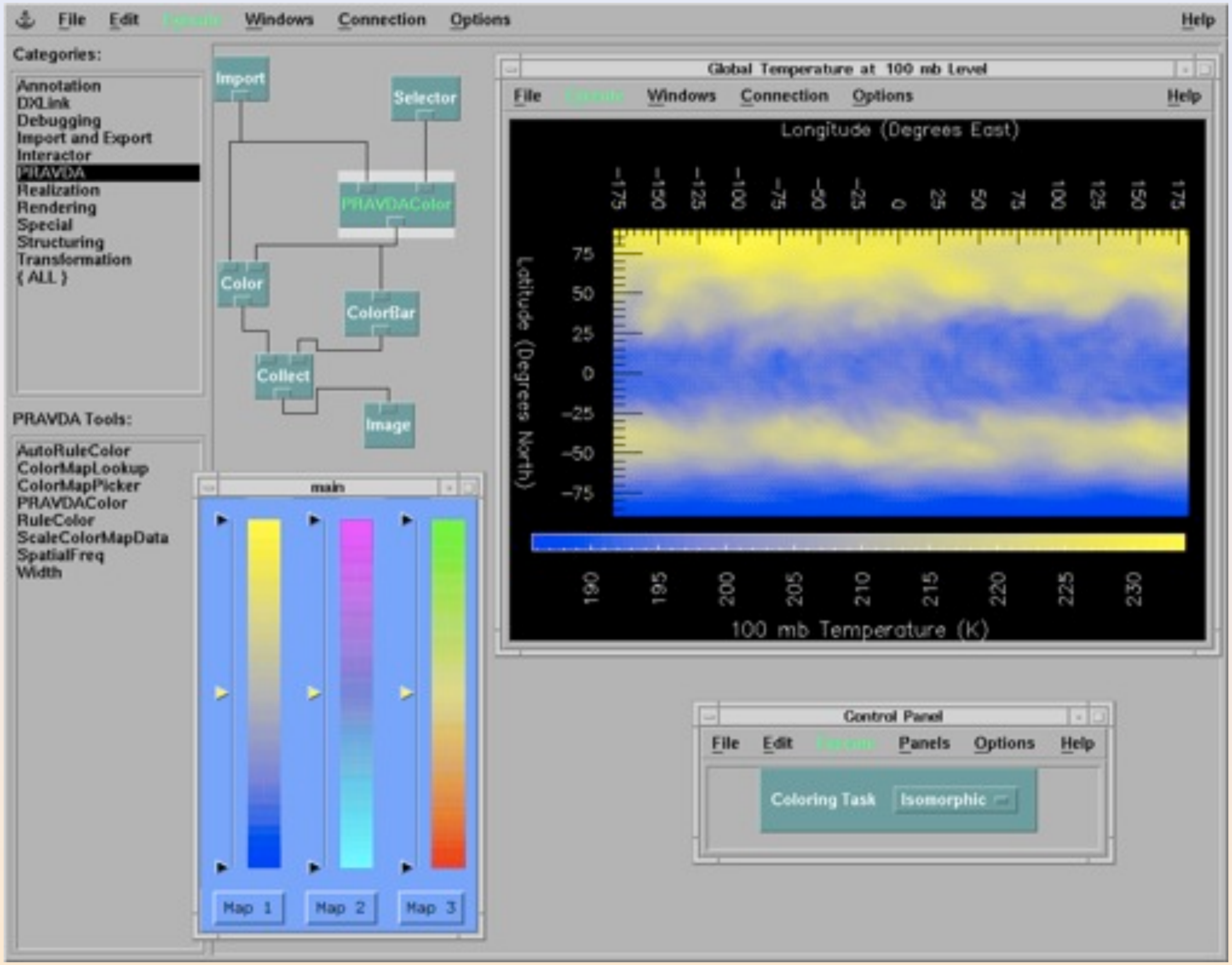
Imposing error analysis methods here would greatly help validate the accuracy of the visualization.

“T

Anderson, Heitmann, Habib, et al

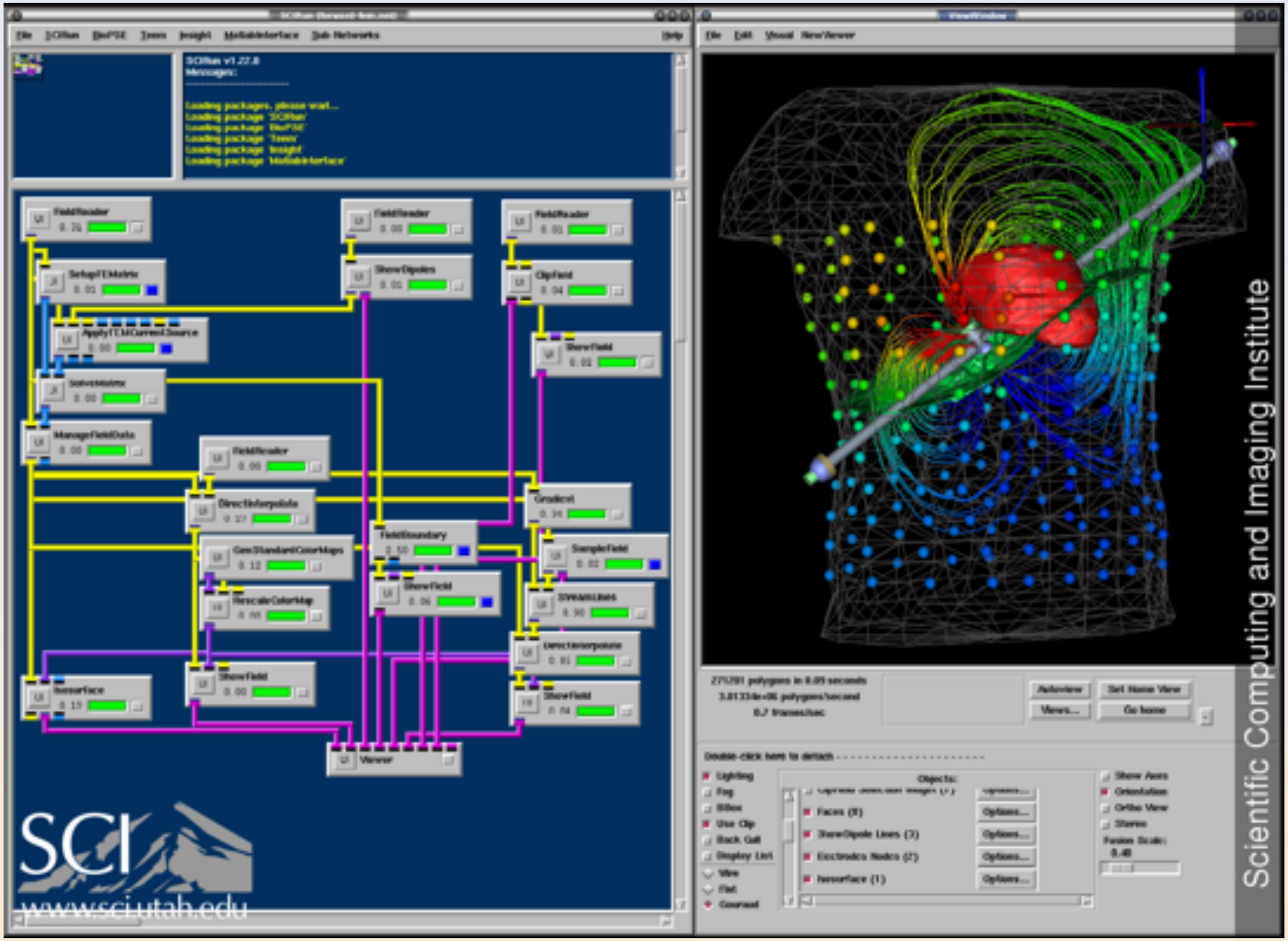
Influential VIS tools

IBM OpenDX



Coloring Task Isomorphic

SCIRun

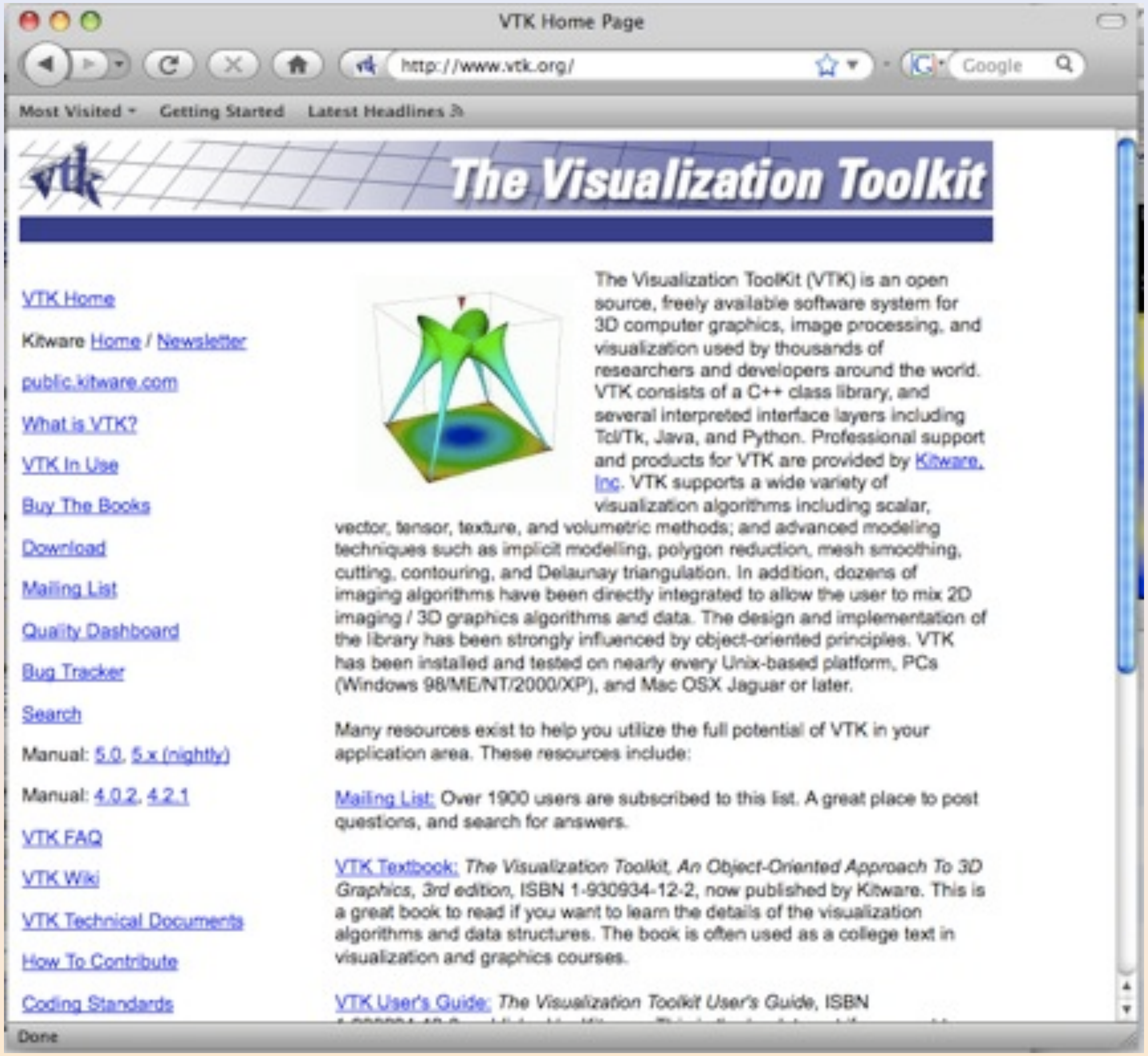


SCIRun v1.22.0
Messages:

Loading packages, please wait...
Loading package 'SCIRun'
Loading package 'BioPSE'
Loading package 'Teem'
Loading package 'Insight'
Loading package 'MatlabInterface'

271281 polygons in 0.09 seconds
3.01334e+06 polygons/second
0.7 frames/sec

VTK



The Visualization Toolkit

- VTK Home
- Kitware Home / Newsletter
- public.kitware.com
- What is VTK?
- VTK In Use
- Buy The Books
- Download
- Mailing List
- Quality Dashboard
- Bug Tracker
- Search
- Manual: 5.0, 5.x (nightly)
- Manual: 4.0.2, 4.2.1
- VTK FAQ
- VTK Wiki
- VTK Technical Documents
- How To Contribute
- Coding Standards

The Visualization ToolKit (VTK) is an open source, freely available software system for 3D computer graphics, image processing, and visualization used by thousands of researchers and developers around the world. VTK consists of a C++ class library, and several interpreted interface layers including Tcl/Tk, Java, and Python. Professional support and products for VTK are provided by Kitware, Inc. VTK supports a wide variety of visualization algorithms including scalar, vector, tensor, texture, and volumetric methods; and advanced modeling techniques such as implicit modelling, polygon reduction, mesh smoothing, cutting, contouring, and Delaunay triangulation. In addition, dozens of imaging algorithms have been directly integrated to allow the user to mix 2D imaging / 3D graphics algorithms and data. The design and implementation of the library has been strongly influenced by object-oriented principles. VTK has been installed and tested on nearly every Unix-based platform, PCs (Windows 98/ME/NT/2000/XP), and Mac OSX Jaguar or later.

Many resources exist to help you utilize the full potential of VTK in your application area. These resources include:

Mailing List: Over 1900 users are subscribed to this list. A great place to post questions, and search for answers.

VTK Textbook: *The Visualization Toolkit, An Object-Oriented Approach To 3D Graphics, 3rd edition*, ISBN 1-930934-12-2, now published by Kitware. This is a great book to read if you want to learn the details of the visualization algorithms and data structures. The book is often used as a college text in visualization and graphics courses.

VTK User's Guide: *The Visualization Toolkit User's Guide*, ISBN

Digression: Workflows

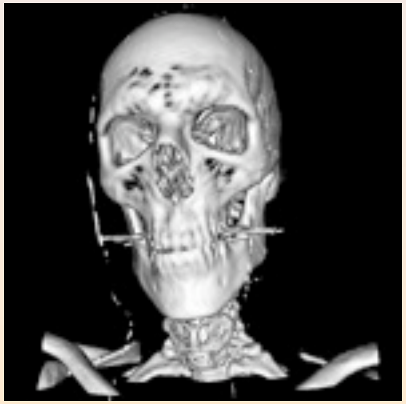
Scientific Workflows and Dataflows

- Dataflows are directed graphs describing a computational task
  - Vertices = modules = processing steps + **parameters**
  - Edges = connections between output and input ports
  - Execution order determined by flow of data from output to input ports

Input:

Head.120.iso

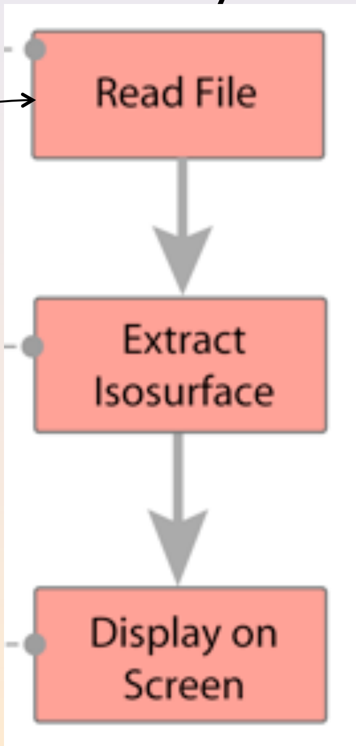
Output:



Scientific Workflows and Dataflows

- Dataflows are directed graphs describing a computational task
  - Vertices = modules = processing steps + **parameters**
  - Edges = connections between output and input ports
  - Execution order determined by flow of data from output to input ports

Input:
Head.120.iso



Isosurface ***value=57***

Output:

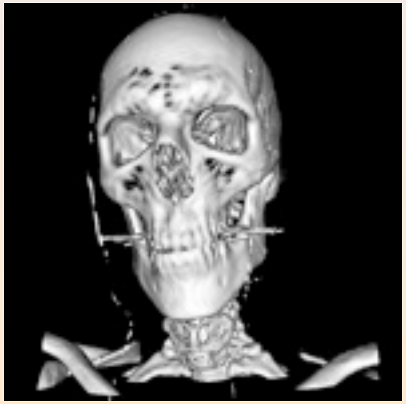
Scientific Workflows and Dataflows

- Dataflows are directed graphs describing a computational task
  - Vertices = modules = processing steps + **parameters**
  - Edges = connections between output and input ports
  - Execution order determined by flow of data from output to input ports

Input:

`Head.120.iso`

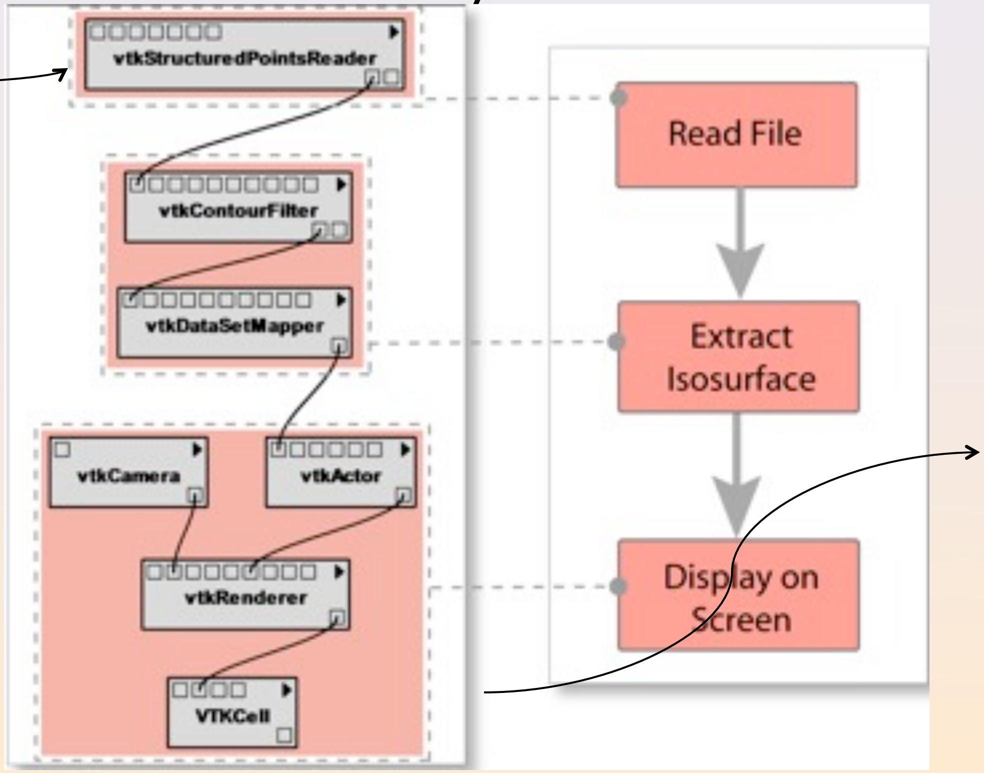
Output:



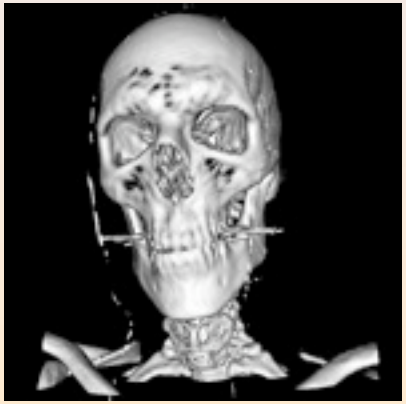
Scientific Workflows and Dataflows

- Dataflows are directed graphs describing a computational task
  - Vertices = modules = processing steps + **parameters**
  - Edges = connections between output and input ports
  - Execution order determined by flow of data from output to input ports

Input:
Head.120.iso

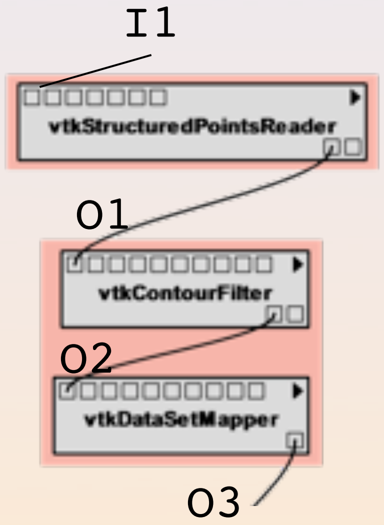


Output:

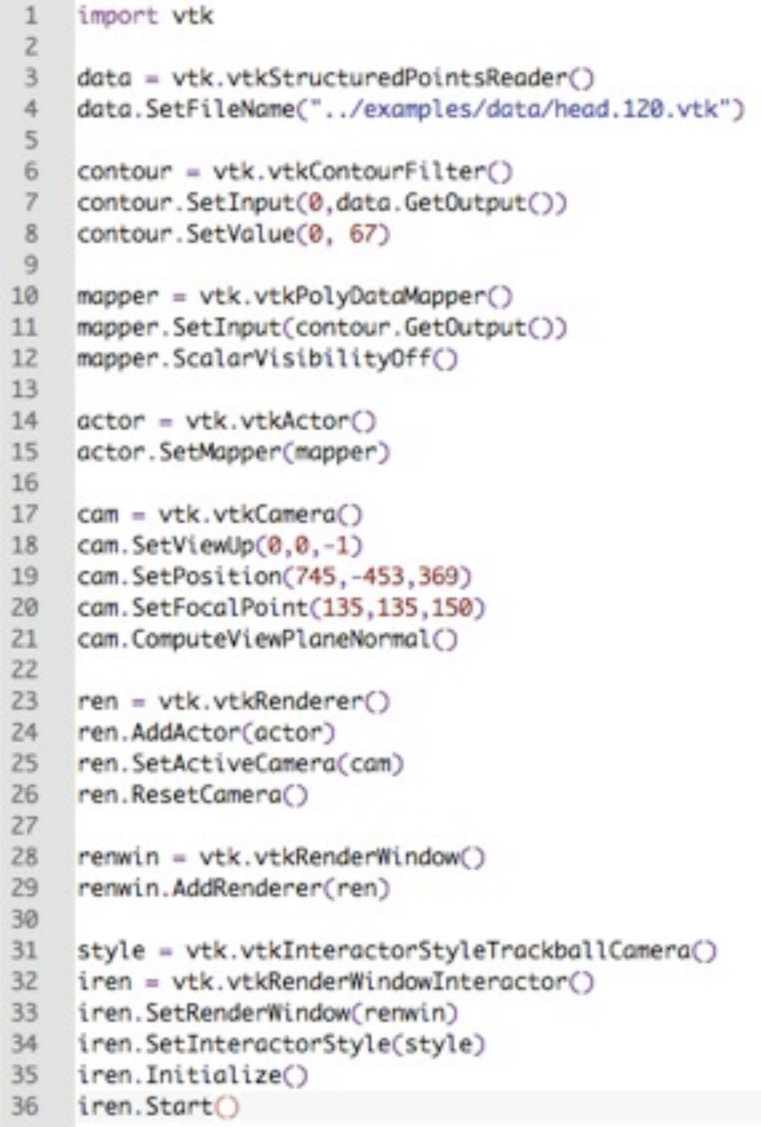


Scientific Workflows and Dataflows

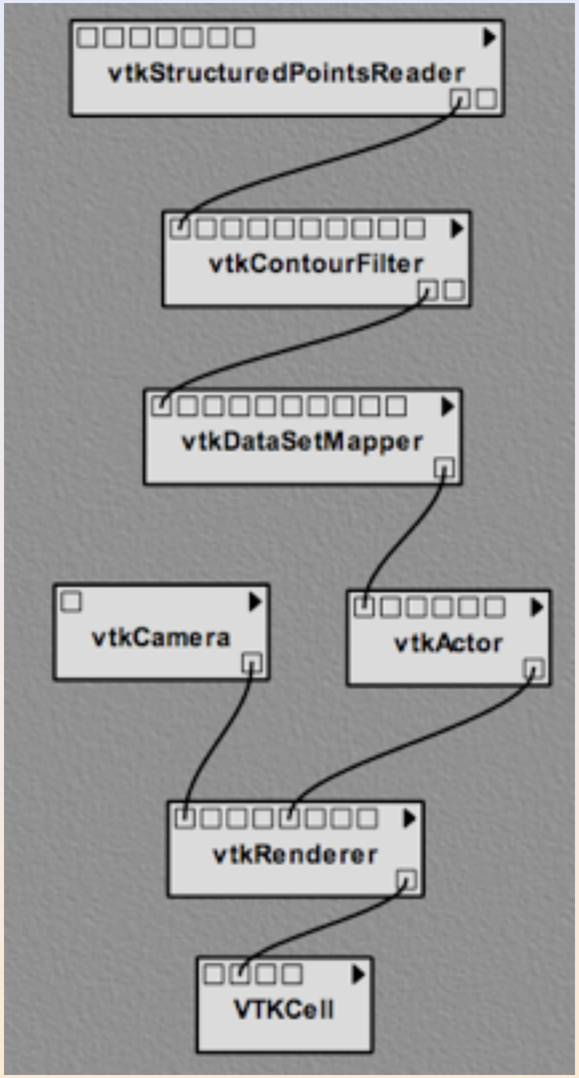
- A directed graph describing a computational task
  - Vertices = modules = processing steps + parameters
  - Edges = connections between output and input ports
  - Execution order determined by flow of data from output to input ports
- No state or side effects: Outputs are a *function* of the inputs
- Simple programming model
  - Good match for visual programming interfaces
  - Widely used: adopted by most scientific workflow and visualization systems
  - Easy to optimize and parallelize



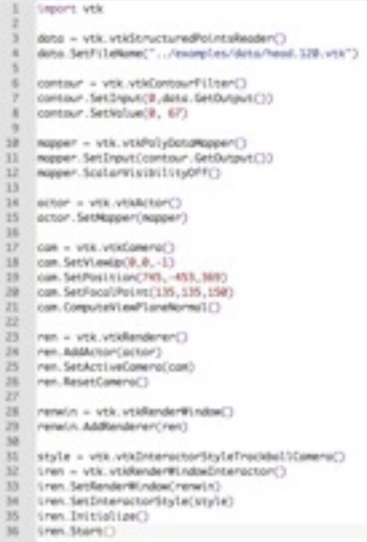
Workflows and Computer Programs



```
import vtk

data = vtk.vtkStructuredPointsReader()
data.SetFileName("../examples/data/head.120.vtk")

contour = vtk.vtkContourFilter()
contour.SetInput(0,data.GetOutput())
contour.SetValue(0, 67)

mapper = vtk.vtkPolyDataMapper()
mapper.SetInput(contour.GetOutput())
mapper.ScalarVisibilityOff()

actor = vtk.vtkActor()
actor.SetMapper(mapper)

cam = vtk.vtkCamera()
cam.SetViewUp(0,0,-1)
cam.SetPosition(745,-453,369)
cam.SetFocalPoint(135,135,150)
cam.ComputeViewPlaneNormal()

ren = vtk.vtkRenderer()
ren.AddActor(actor)
ren.SetActiveCamera(cam)
ren.ResetCamera()

renwin = vtk.vtkRenderWindow()
renwin.AddRenderer(ren)

style = vtk.vtkInteractorStyleTrackballCamera()
iren = vtk.vtkRenderWindowInteractor()
iren.SetRenderWindow(renwin)
iren.SetInteractorStyle(style)
iren.Initialize()
iren.Start()
```

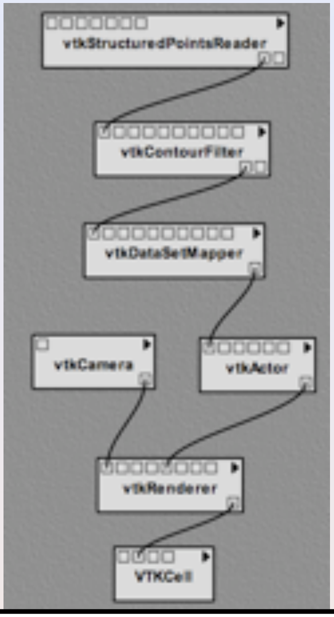


Workflows and Computer Programs

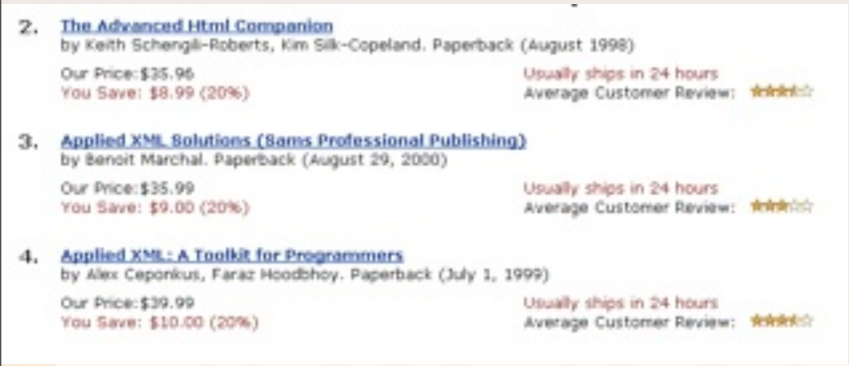


```
import vtk

data = vtk.vtkStructuredPointsReader()
data.SetFileName("../examples/data/head.120.vtk")

contour = vtk.vtkContourFilter()
contour.SetInput(0,data.GetOutput())
contour.SetValue(0, 67)

mapper = vtk.vtkPolyDataMapper()
mapper.SetInput(contour.GetOutput())
mapper.ScalarVisibilityOff()

actor = vtk.vtkActor()
actor.SetMapper(mapper)

cam = vtk.vtkCamera()
cam.SetViewUp(0,0,-1)
cam.SetPosition(745,-453,369)
cam.SetFocalPoint(135,135,150)
cam.ComputeViewPlaneNormal()

ren = vtk.vtkRenderer()
ren.AddActor(actor)
ren.SetActiveCamera(cam)
ren.ResetCamera()

renwin = vtk.vtkRenderWindow()
renwin.AddRenderer(ren)

style = vtk.vtkInteractorStyleTrackballCamera()
iren = vtk.vtkRenderWindowInteractor()
iren.SetRenderWindow(renwin)
iren.SetInteractorStyle(style)
iren.Initialize()
iren.Start()
```

Program → Workflow



Document → Database



2. **The Advanced Html Companion**
by Keith Schengili-Roberts, Kim Silk-Copeland. Paperback (August 1998)
Our Price:$35.96
You Save: $8.99 (20%)
Usually ships in 24 hours
Average Customer Review: ★★★★☆

3. **Applied XML Solutions (Sams Professional Publishing)**
by Benoit Marchal. Paperback (August 29, 2000)
Our Price:$35.99
You Save: $9.00 (20%)
Usually ships in 24 hours
Average Customer Review: ★★★☆☆

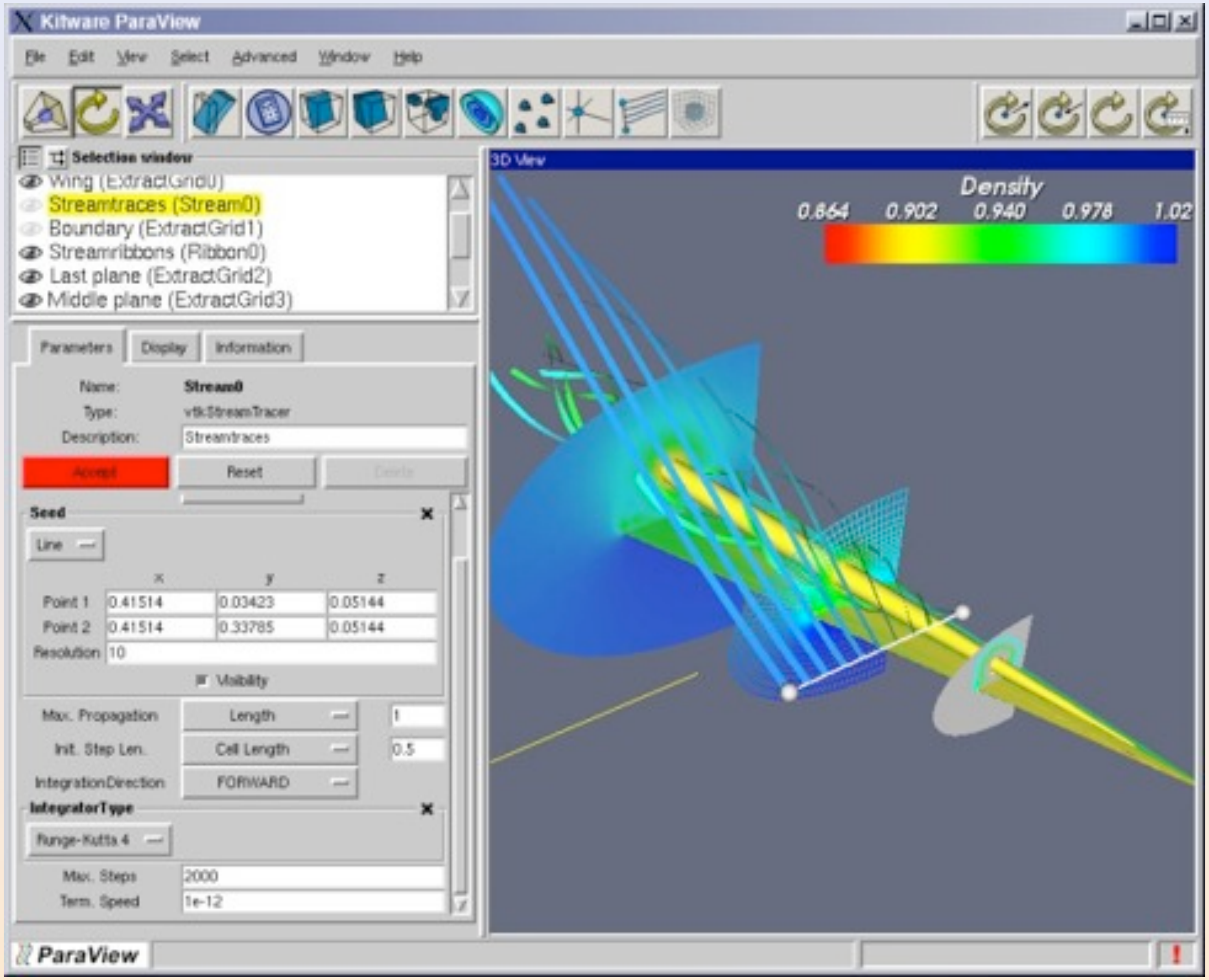
4. **Applied XML: A Toolkit for Programmers**
by Alex Ceponkus, Faraz Hoodbhoy. Paperback (July 1, 1999)
Our Price:$39.99
You Save: $10.00 (20%)
Usually ships in 24 hours
Average Customer Review: ★★★★☆

```
<Book>
<Title>The Advanced Html Companion</Title>
<Author> Keith Schengili-Roberts </Author>
<Author> Kim Silk-Copeland</Author>
<Price> 35.96</Price>…
</Book>
```

A program is to a workflow what an unstructured document is to a (structured) database.

Back to VIS

ParaView



Wing (ExtractGrid0)
Streamtraces (Stream0)
Boundary (ExtractGrid1)
Streamribbons (Ribbon0)
Last plane (ExtractGrid2)
Middle plane (ExtractGrid3)

Name: Stream0
Type: vtkStreamTracer
Description: Streamtraces

Seed

Line

	x	y	z
Point 1	0.41514	0.03423	0.05144
Point 2	0.41514	0.33785	0.05144

Resolution 10

☒ Visibility

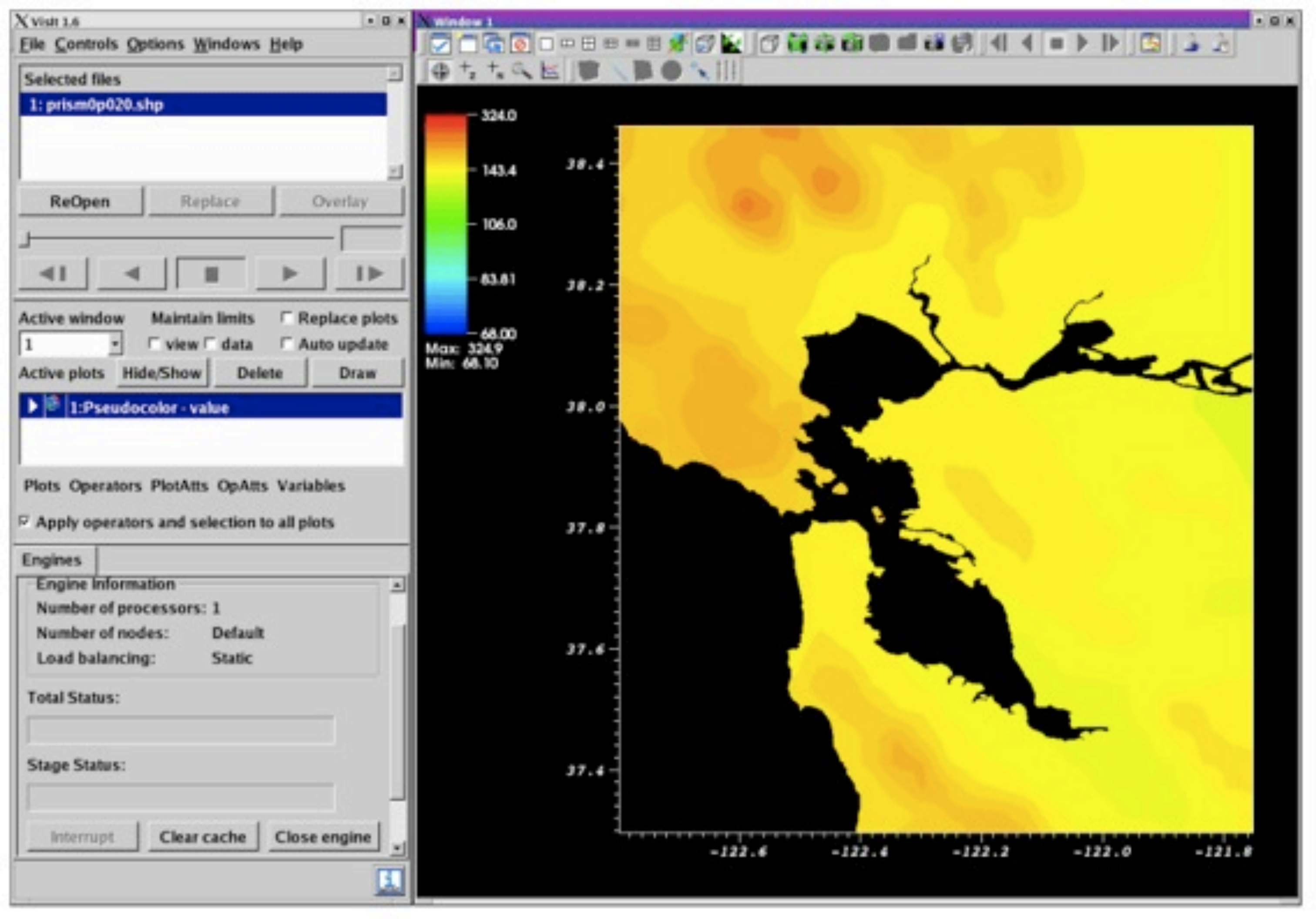
Max. Propagation: Length 1
Init. Step Len.: Cell Length 0.5
IntegrationDirection: FORWARD

IntegratorType

Runge-Kutta 4

Max. Steps 2000
Term. Speed 1e-12

VisIt



1: prism0p020.shp

Active window: 1

1:Pseudocolor - value

☑ Apply operators and selection to all plots

Engine Information

Number of processors: 1

Number of nodes: Default

Load balancing: Static

Total Status:

Stage Status:

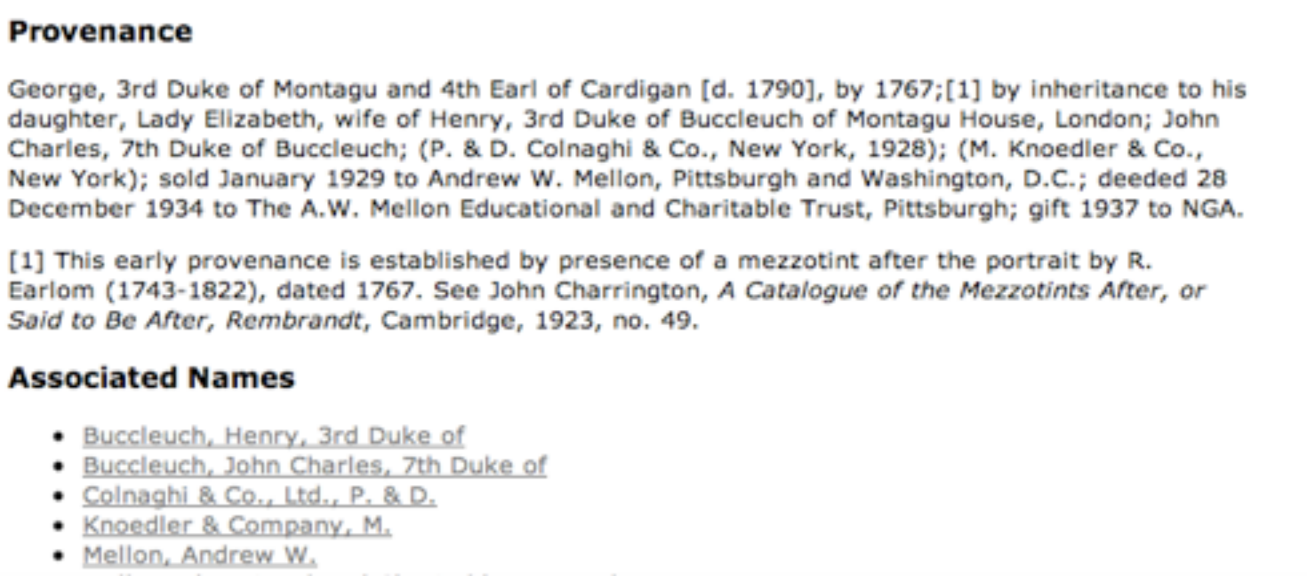
Max: 324.9

Min: 68.10

VisTrails Project

Provenance in Art

Rembrandt van Rijn
Self-Portrait, 1659
Andrew W. Mellon Collection
1937.1.72



Provenance

George, 3rd Duke of Montagu and 4th Earl of Cardigan [d. 1790], by 1767;[1] by inheritance to his daughter, Lady Elizabeth, wife of Henry, 3rd Duke of Buccleuch of Montagu House, London; John Charles, 7th Duke of Buccleuch; (P. & D. Colnaghi & Co., New York, 1928); (M. Knoedler & Co., New York); sold January 1929 to Andrew W. Mellon, Pittsburgh and Washington, D.C.; deeded 28 December 1934 to The A.W. Mellon Educational and Charitable Trust, Pittsburgh; gift 1937 to NGA.

[1] This early provenance is established by presence of a mezzotint after the portrait by R. Earlom (1743-1822), dated 1767. See John Charrington, *A Catalogue of the Mezzotints After, or Said to Be After, Rembrandt*, Cambridge, 1923, no. 49.

Associated Names

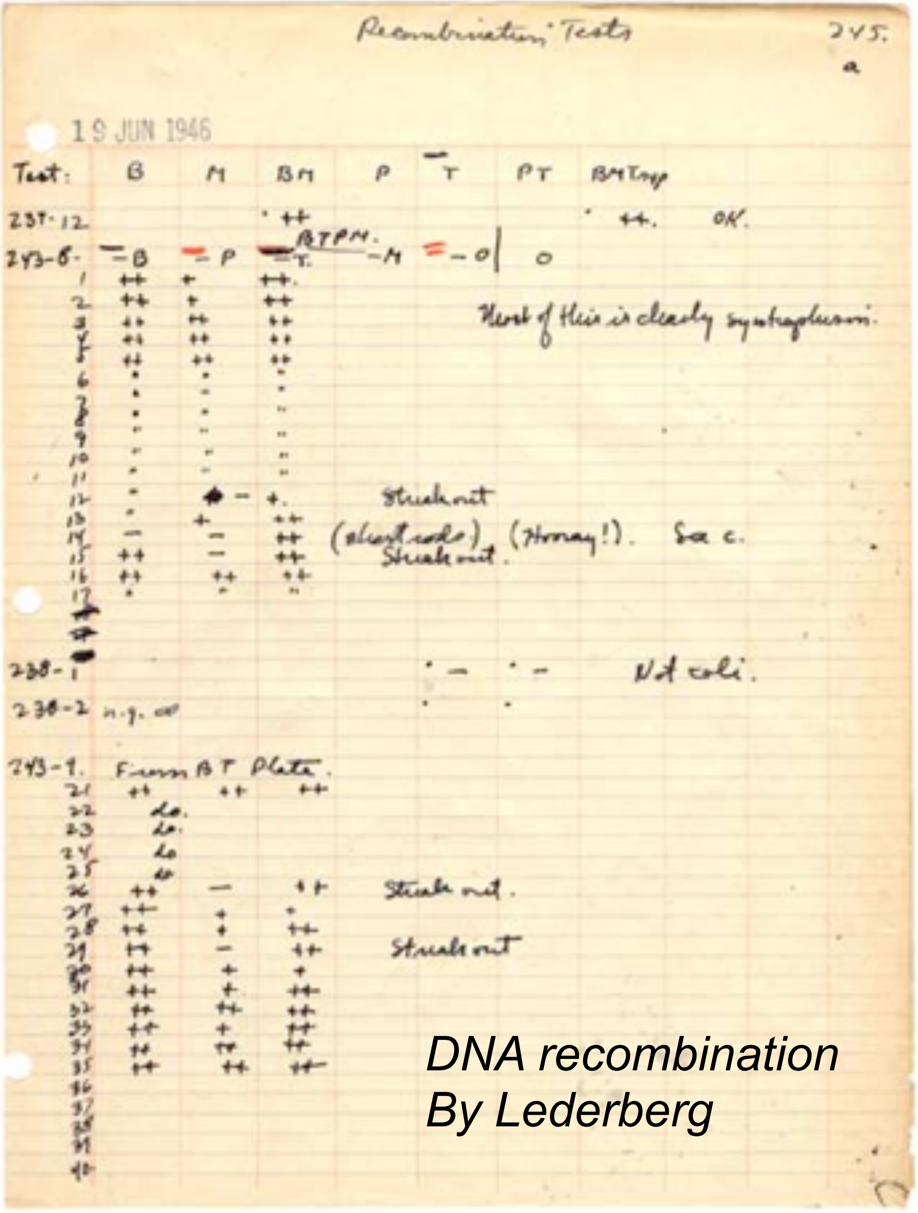
- Buccleuch, Henry, 3rd Duke of
- Buccleuch, John Charles, 7th Duke of
- Colnaghi & Co., Ltd., P. & D.
- Knoedler & Company, M.
- Mellon, Andrew W.

Provenance in Science

- *Provenance is as (or more!) important as the result*
- Not a new issue
- Lab notebooks have

been used for a long time

- What is new?
  - Large volumes of data
  - Complex analyses
- Writing notes is no longer an option…
- Computational provenance



Recombination Tests

245.
a

19 JUN 1946

Test:	B	M	BM	P	T̄	PT	BMTnp
235-12			· ++				· ++. OK.
243-6-	–B	–P	–T. BTPM.	–M	=–0	0	
1	++	+	++.				
2	++	+	++		Most of this is clearly syntrophism.		
3	++	++	++				
4	++	++	++				
5	++	++	++				
6	·	·	·				
7	·	·	·				
8	·	·	··				
9	··	··	··				
10	·	·	··				
11	·	·	··				
12	·	♦ –	+.	Streak out			
13	·	+	++				
14	–	–	++	(short code)	(Hooray!).	See c.	
15	++	–	++	Streak out.			
16	++	++	++				
17	·	·	··				
238-1				· –	· –	Not coli.	
238-2	n.g. [illegible]			·	·		
243-1.	From BT Plate.						
21	++	++	++				
22	do.						
23	do.						
24	do						
25	do						
26	++	–	++	Streak out.			
27	++	+	·				
28	++	+	++				
29	++	–	++	Streak out			
30	++	+	+				
31	++	+	++				
32	++	++	++				
33	++	+	++				
34	++	++	++				
35	++	++	++				
36							
37							
38							
39							
40							

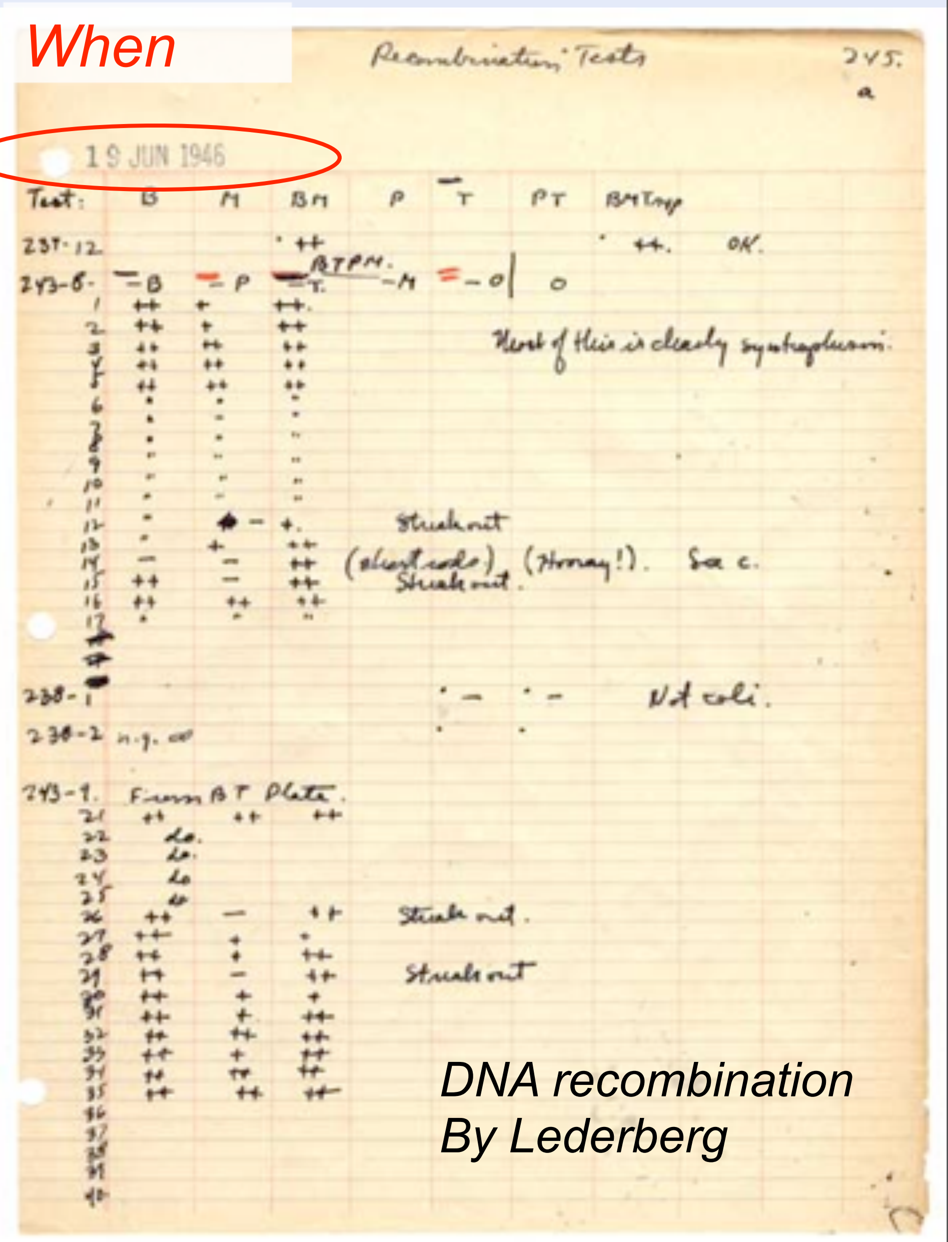
DNA recombination
By Lederberg

Provenance in Science

- *Provenance is as (or more!) important as the result*
- Not a new issue
- Lab notebooks have

been used for a long time

- What is new?
  - Large volumes of data
  - Complex analyses
- Writing notes is no longer an option…
- Computational provenance



When

Recombination Tests

245.

a

19 JUN 1946

Test:	B	M	BM	P	T	PT	BMTaxy	
235-12			· ++				· ++.	OK.
243-6-	=B	=P	=T. BTPM.	-M	=-0|	0		
1	++	+	++.					
2	++	+	++			Most of this is clearly syntrophism.		
3	++	++	++					
4	++	++	++					
5	++	++	++					
6	·	·	·					
7	·	·	·					
8	·	·	··					
9	·	··	··					
10	·	·	··					
11	·	·	··					
12	·	◆ –	+.	Streak out				
13	·	+	++					
14	–	–	++	(short rods)	(Hooray!).		See c.	
15	++	–	++	Streak out.				
16	++	++	++					
17	·	·	··					
238-1				· –		· –	Not coli.	
238-2	n.g.			·		·		
243-9.	From BT Plate.							
21	++	++	++					
22	do.							
23	do.							
24	do							
25	do							
26	++	–	++	Streak out.				
27	++	+	+					
28	++	+	++					
29	++	–	++	Streak out				
30	++	+	+					
31	++	+	++					
32	++	++	++					
33	++	+	++					
34	++	++	++					
35	++	++	++					
36								
37								
38								
39								
40								

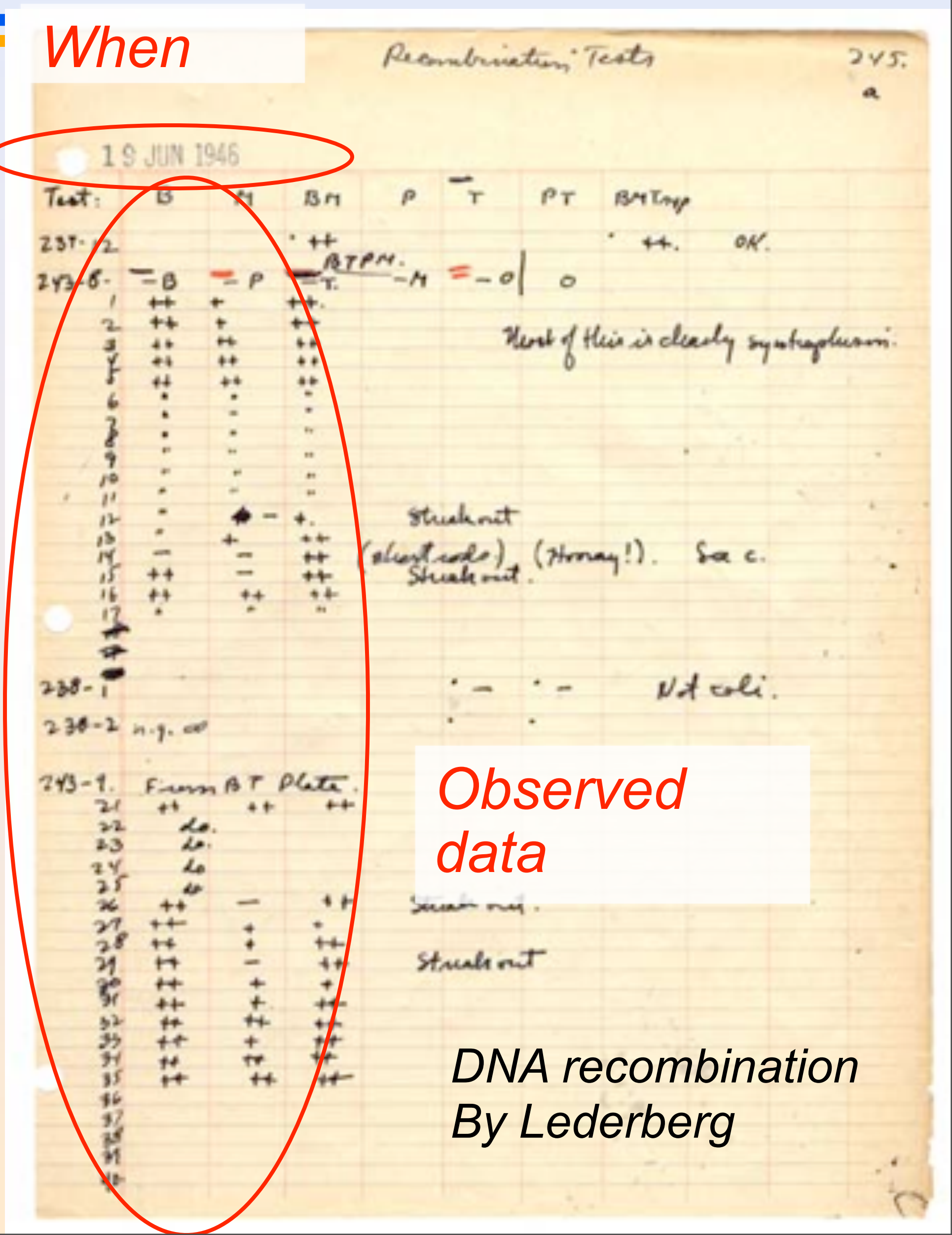
DNA recombination
By Lederberg

Provenance in Science

- *Provenance is as (or more!) important as the result*
- Not a new issue
- Lab notebooks have

been used for a long time

- What is new?
  - Large volumes of data
  - Complex analyses
- Writing notes is no longer an option…
- Computational provenance


Recombination Tests
19 JUN 1946
Test:
B
M
BM
P
T
PT
OK.
Streak out
N.A coli.
From BT Plate

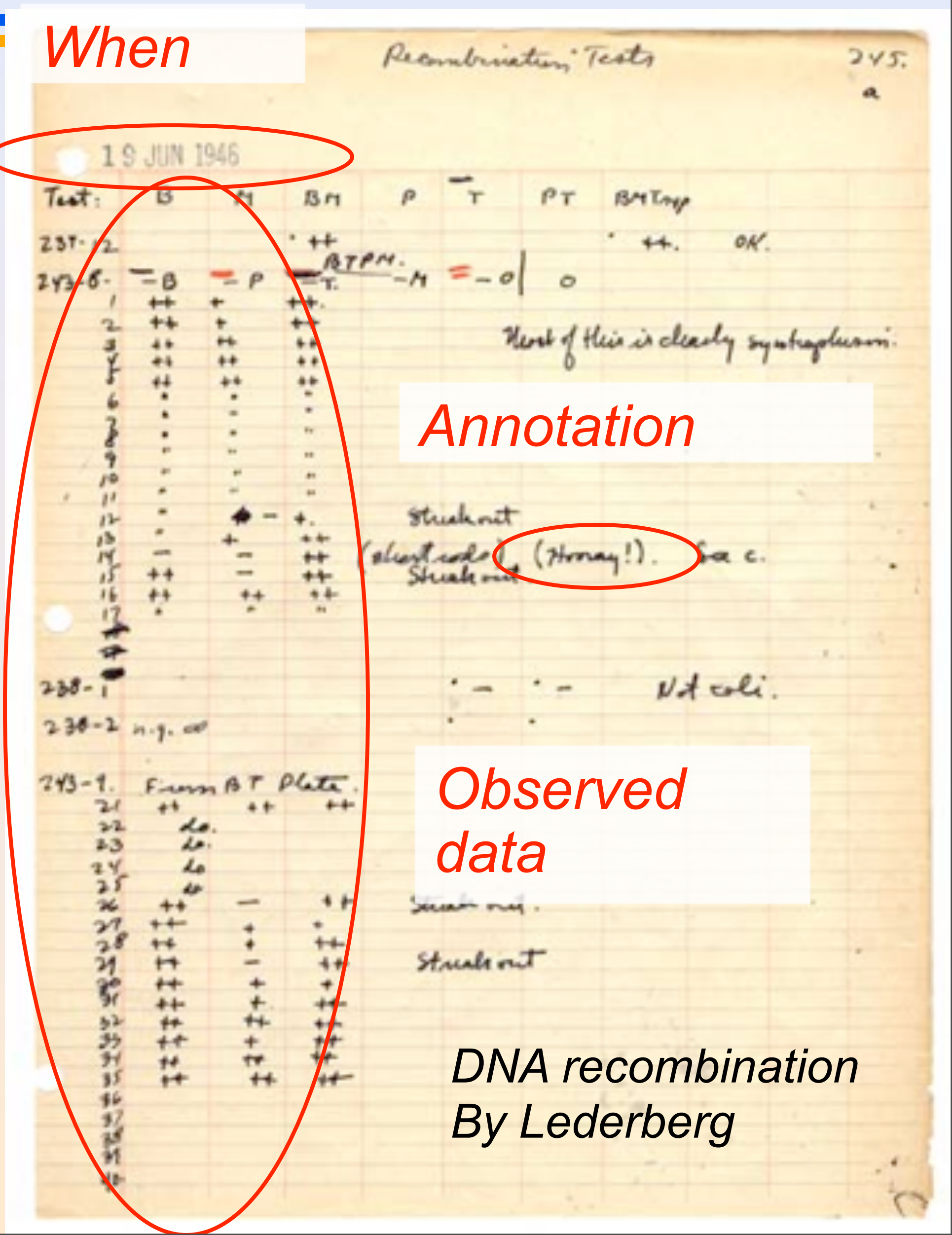
When

Observed data

DNA recombination By Lederberg

Provenance in Science

- *Provenance is as (or more!) important as the result*
- Not a new issue
- Lab notebooks have been used for a long time
- What is new?
  - Large volumes of data
  - Complex analyses
- Writing notes is no longer an option…
- Computational provenance



Recombination Tests 245.

19 JUN 1946

Test: B M BM P T PT BMTaxp

Most of this is clearly syntrophism.

Stuck out

(Hooray!)

Nd coli.

From BT Plate

When

Annotation

Observed data

DNA recombination
By Lederberg

Exploration and Workflows

- Workflows have been traditionally used to automate repetitive tasks
- In exploratory tasks, *change is the norm*!
  - Data analysis and exploration are iterative processes

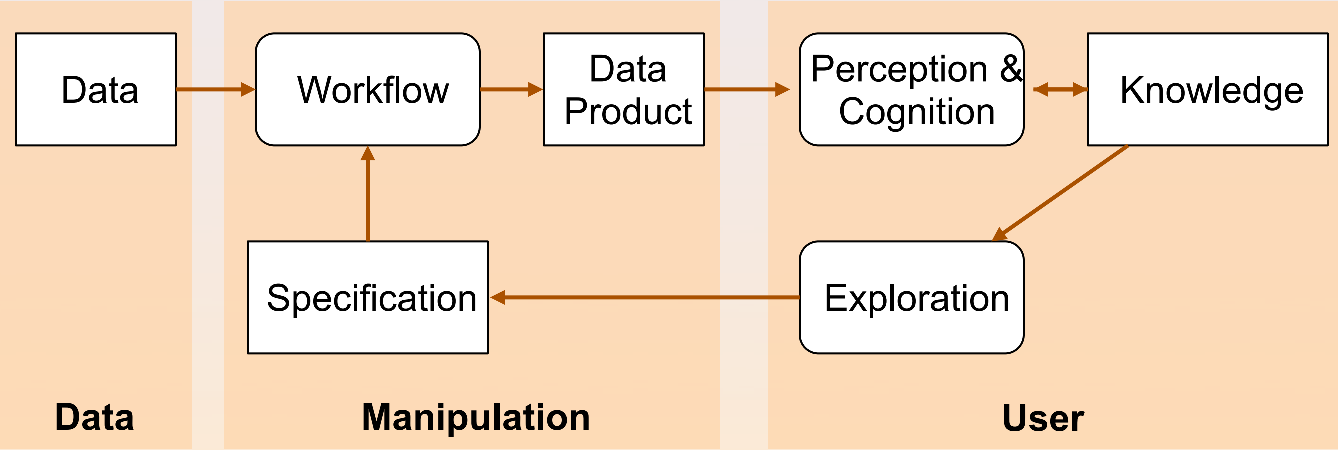


Figure modified from J. van Wijk, IEEE Vis 2005

Exploration and Creativity Support

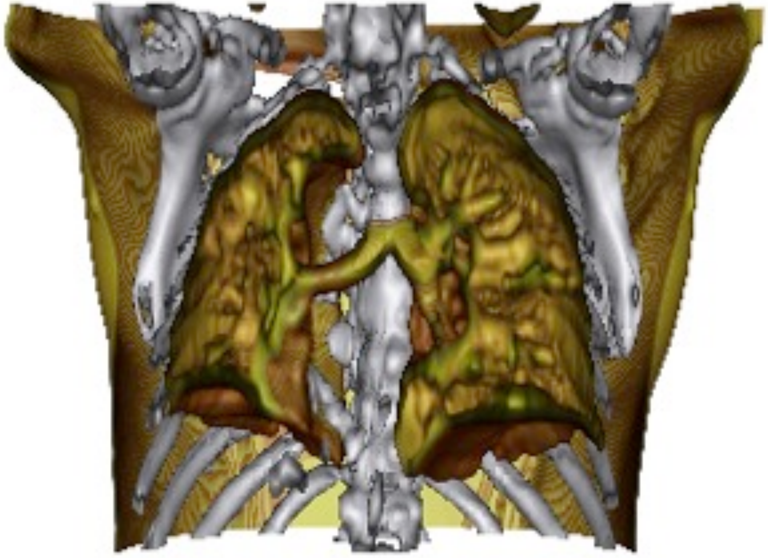
- Reflective reasoning is key in the exploratory processes
- "*Reflective reasoning requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. …the process is slow and laborious*"

  Donald A. Norman
- Need external aids—tools to facilitate this process
  - Creativity support tools [Shneiderman, CACM 2002]
- Need aid from people—collaboration

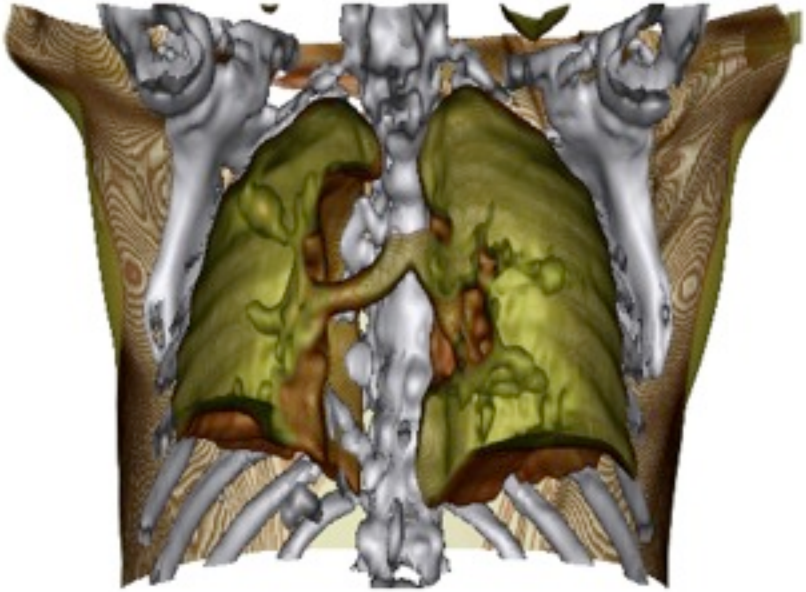
The need for provenance

What’s the difference?

anon4877_base_20060331.jpg



anon4877_lesion_20060401.jpg

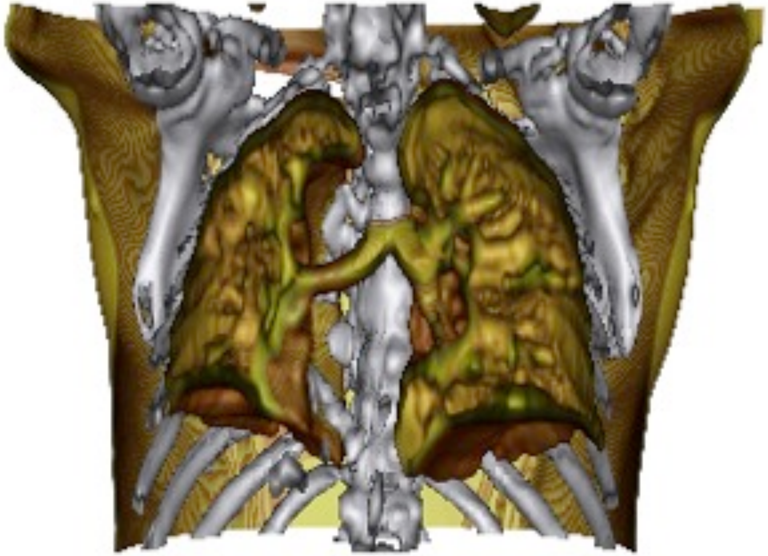


How were these images created?

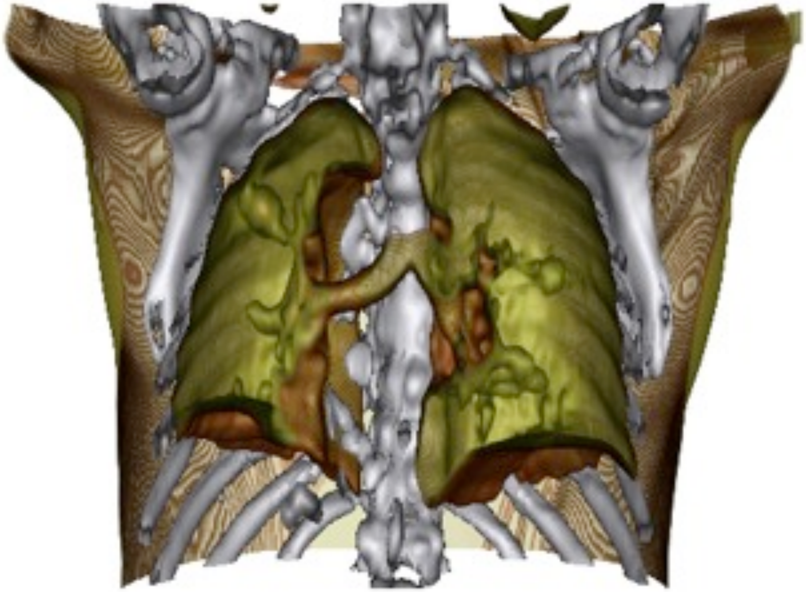
Are they really from the same patient?

The need for provenance

anon4876_base_20060331.jpg



anon4877_lesion_20060401.jpg



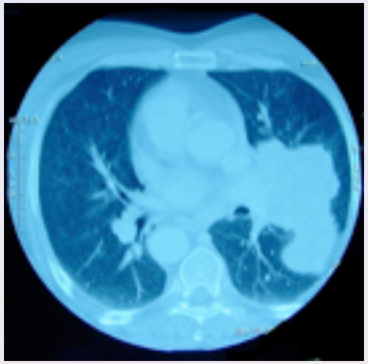
How were these images created?

Are they really from the same patient?

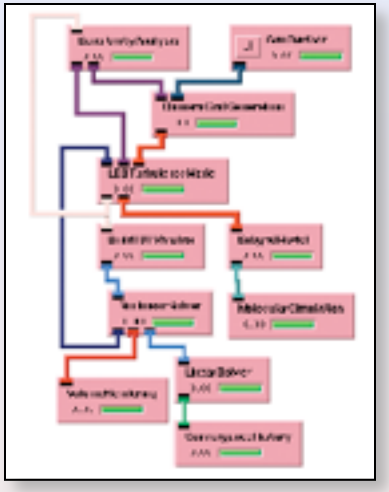
Data Exploration and Workflows

Data Exploration and Workflows

raw data:CT scan

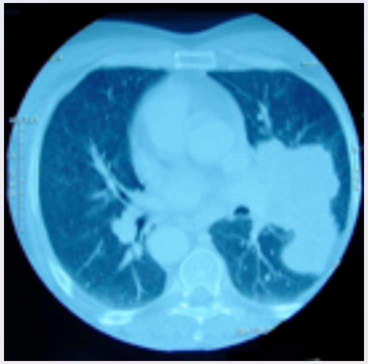


workflow

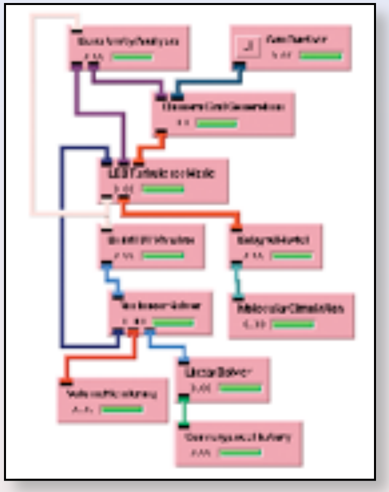


Data Exploration and Workflows

raw data:CT scan

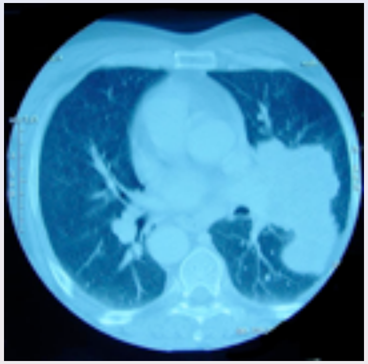


workflow

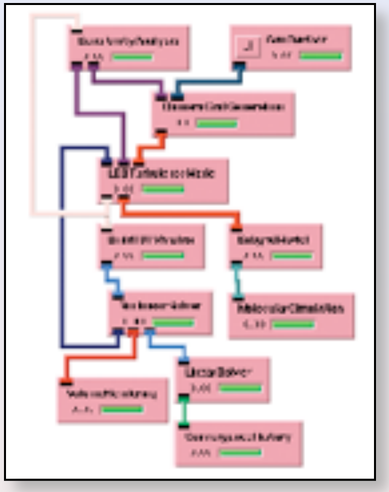


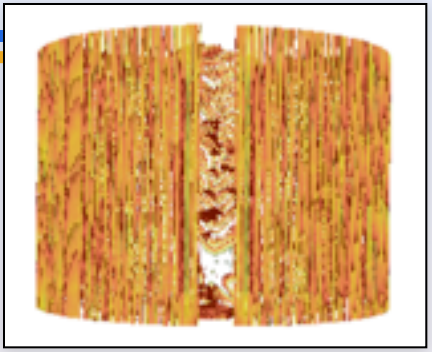
Data Exploration and Workflows

raw data:CT scan



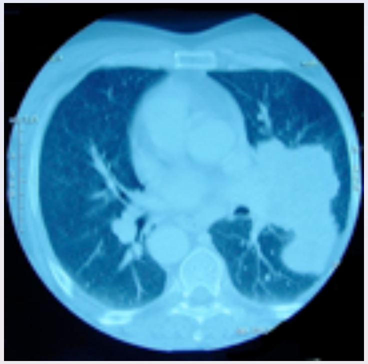
workflow



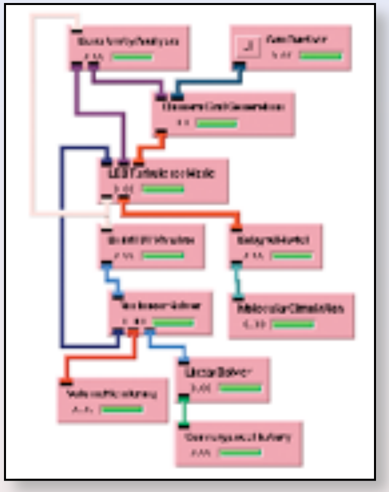


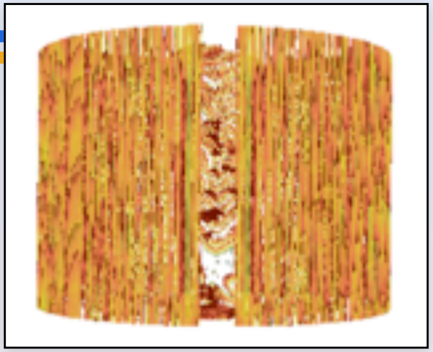
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

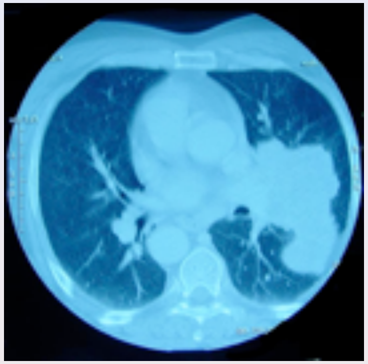
anon4877_voxel_scale_1_zspace_20060331.srn

Notes

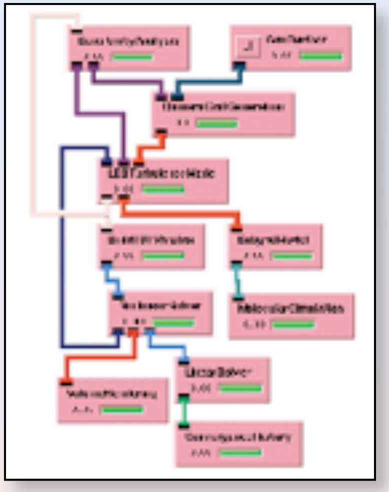
Initial visualization with z-scaling corrected

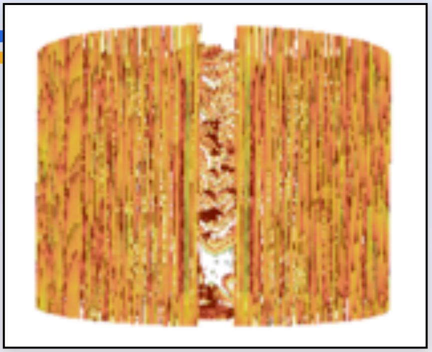
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

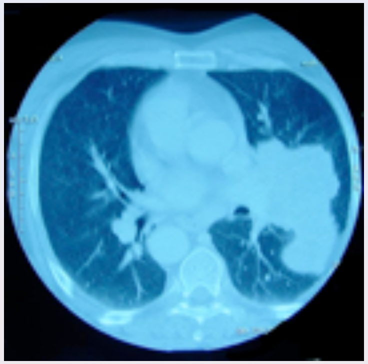
anon4877_voxel_scale_1_zspace_20060331.srn

Notes

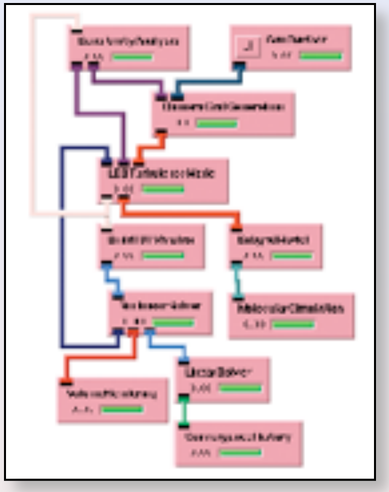
Initial visualization with z-scaling corrected

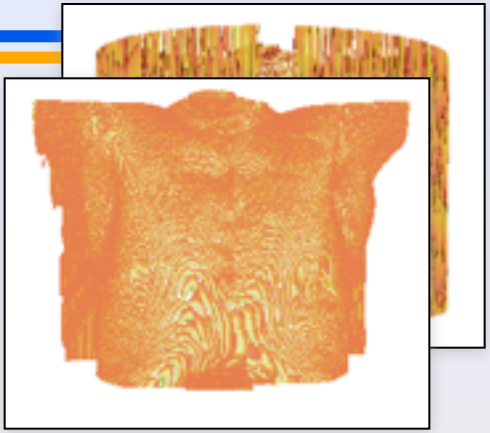
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

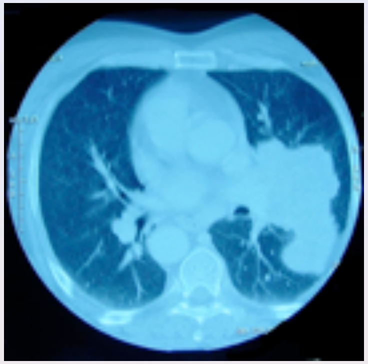
anon4877_voxel_scale_1_zspace_20060331.srn

Notes

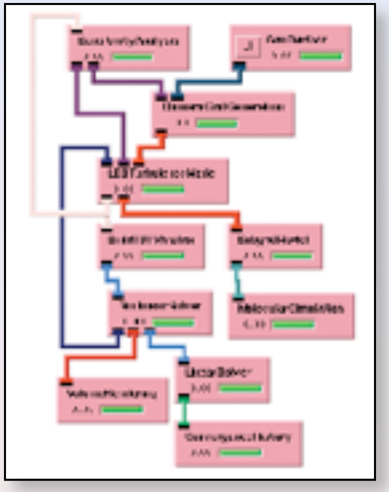
Initial visualization with z-scaling corrected

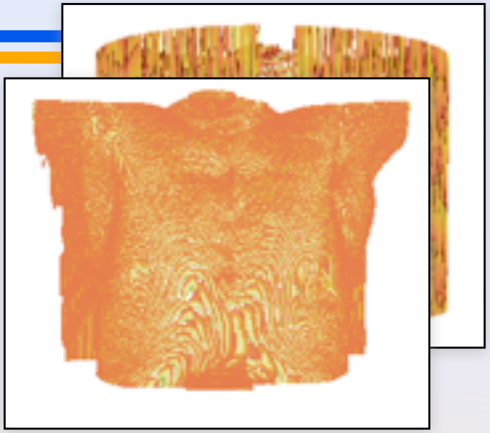
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

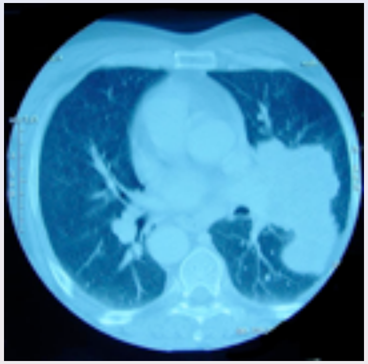
Notes

Initial visualization wi

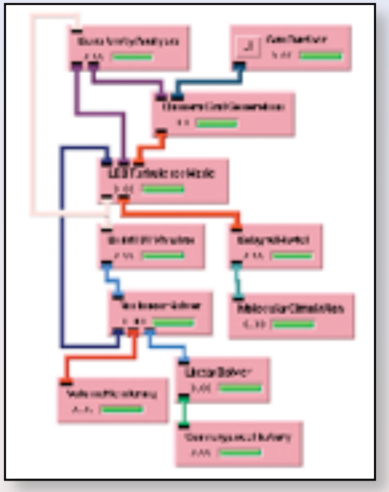
Added texture and shading

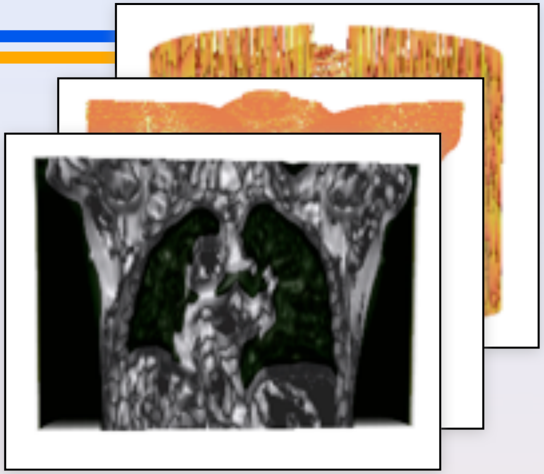
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

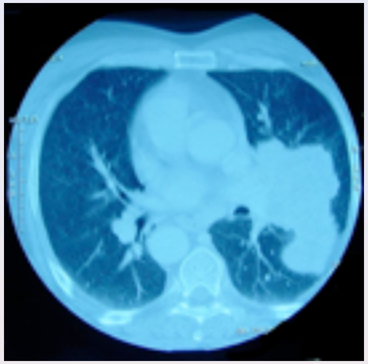
Notes

Initial visualization wi

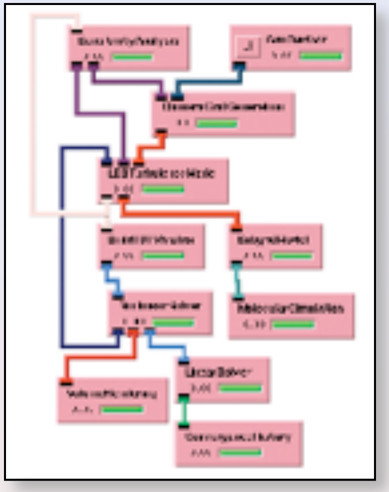
Added texture and shading

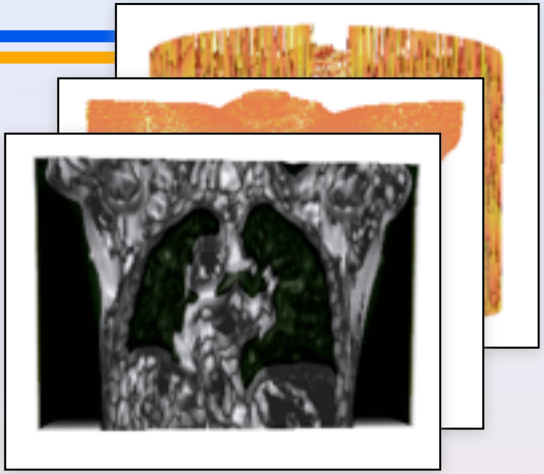
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

anon4877_textureshading_plane0_20060331.srn

Notes

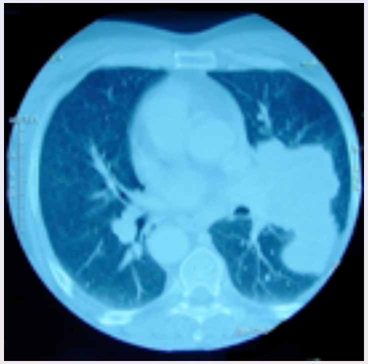
Initial visualization wi

Added texture an

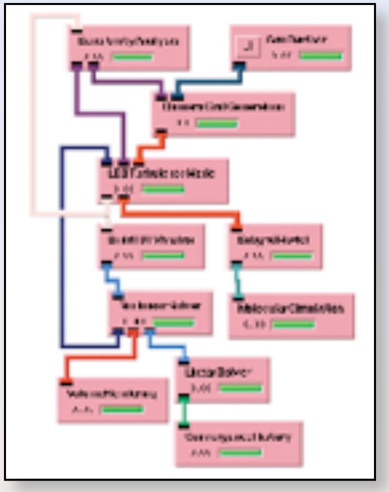
Added plane to visualize internal structure

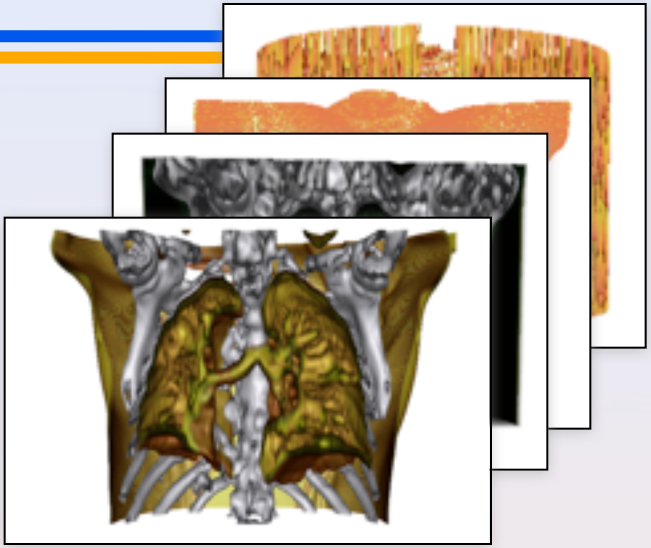
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

anon4877_textureshading_plane0_20060331.srn

Notes

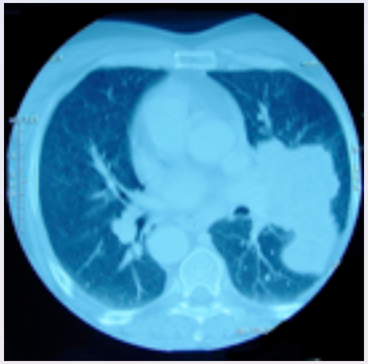
Initial visualization wi

Added texture an

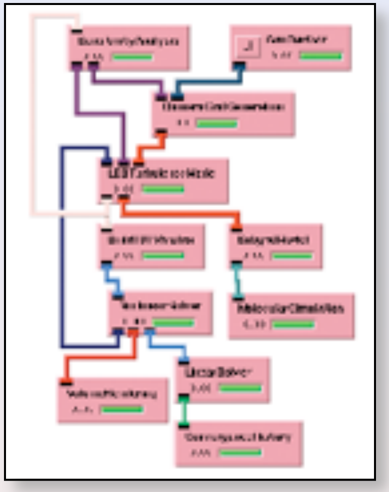
Added plane to visualize internal structure

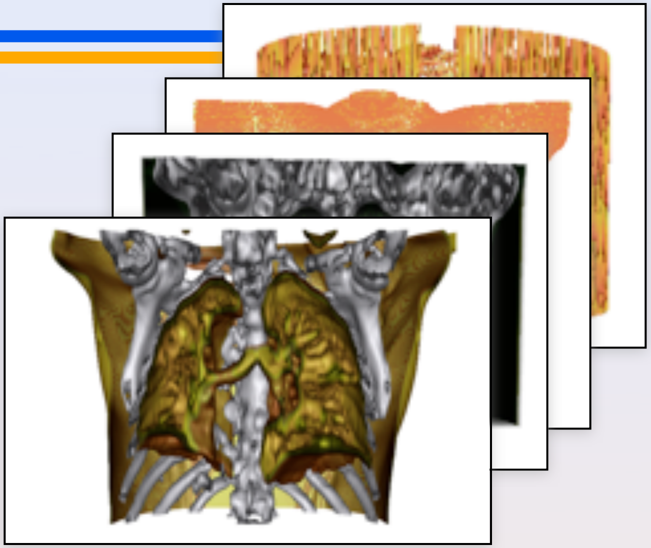
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

anon4877_textureshading_plane0_20060331.srn

anon4877_goodxferfunction_20060331.srn

Notes

Initial visualization wi

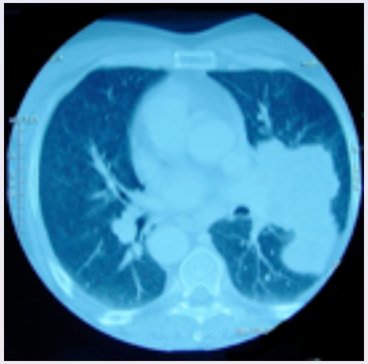
Added texture an

Added plane to v i s

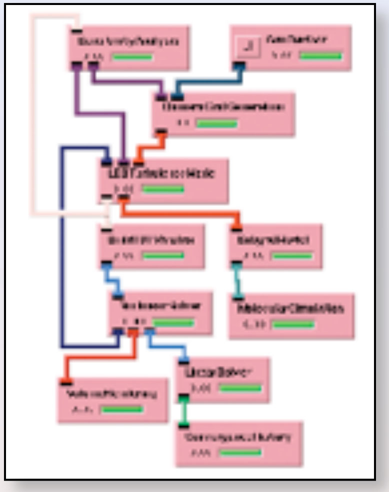
Found good transfer function

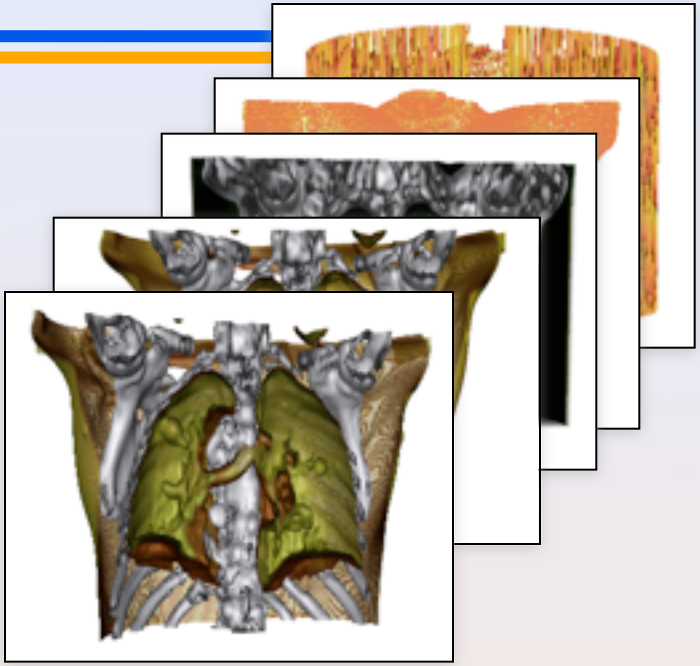
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

anon4877_textureshading_plane0_20060331.srn

anon4877_goodxferfunction_20060331.srn

Notes

Initial visualization wi

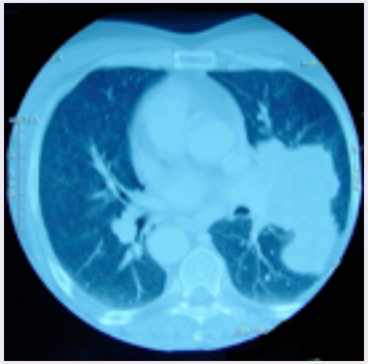
Added texture an

Added plane to v i s

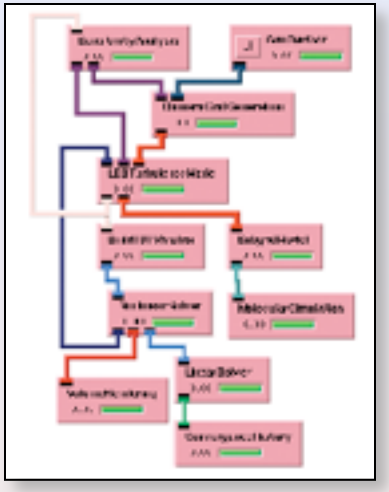
Found good transfer function

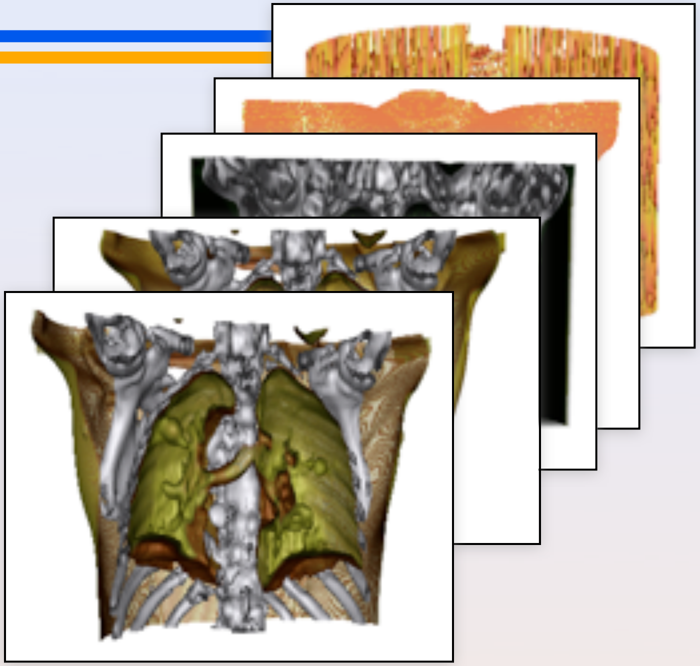
Data Exploration and Workflows

raw data:CT scan



workflow





Files (workflow specifications)

anon4877_voxel_scale_1_zspace_20060331.srn

anon4877_textureshading_20060331.srn

anon4877_textureshading_plane0_20060331.srn

anon4877_goodxferfunction_20060331.srn

anon4877_lesion_20060331.srn

Notes

Initial visualization wi

Added texture an

Added plane to v i s

Found good t f

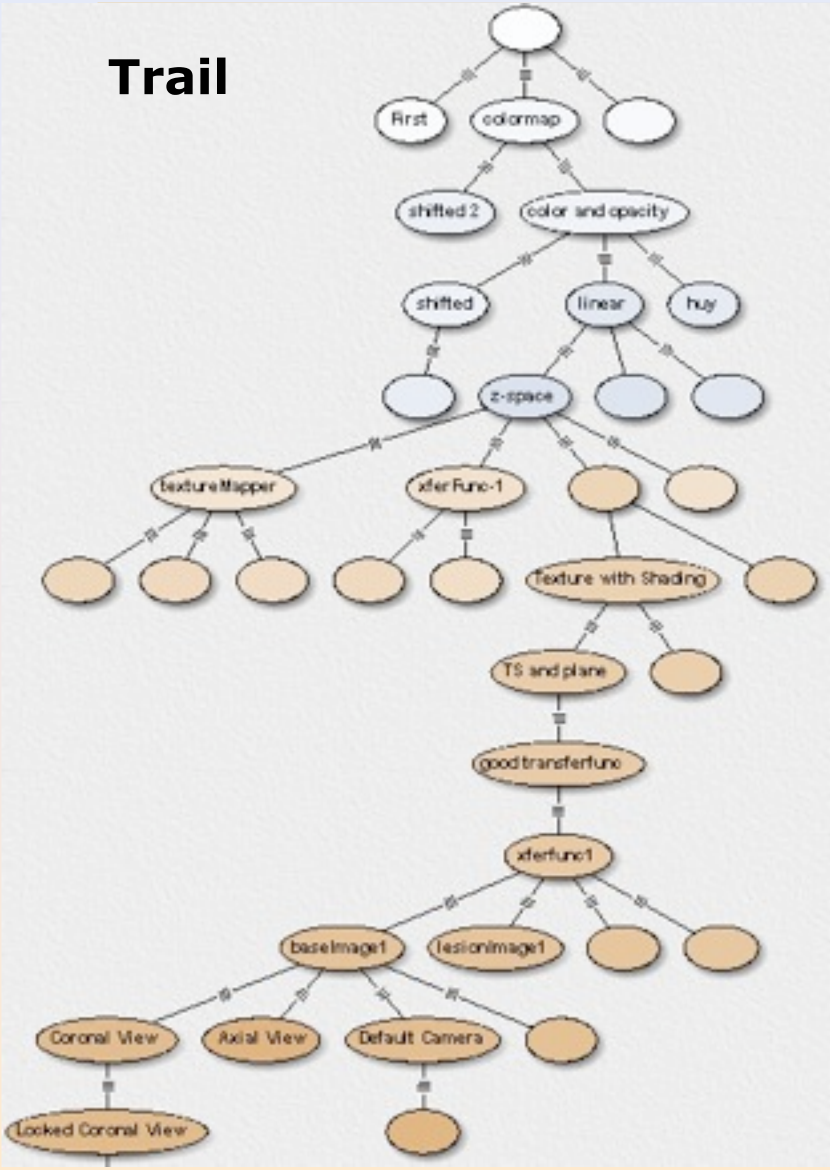
Identified lesion tissue

VisTrails: Managing Exploration

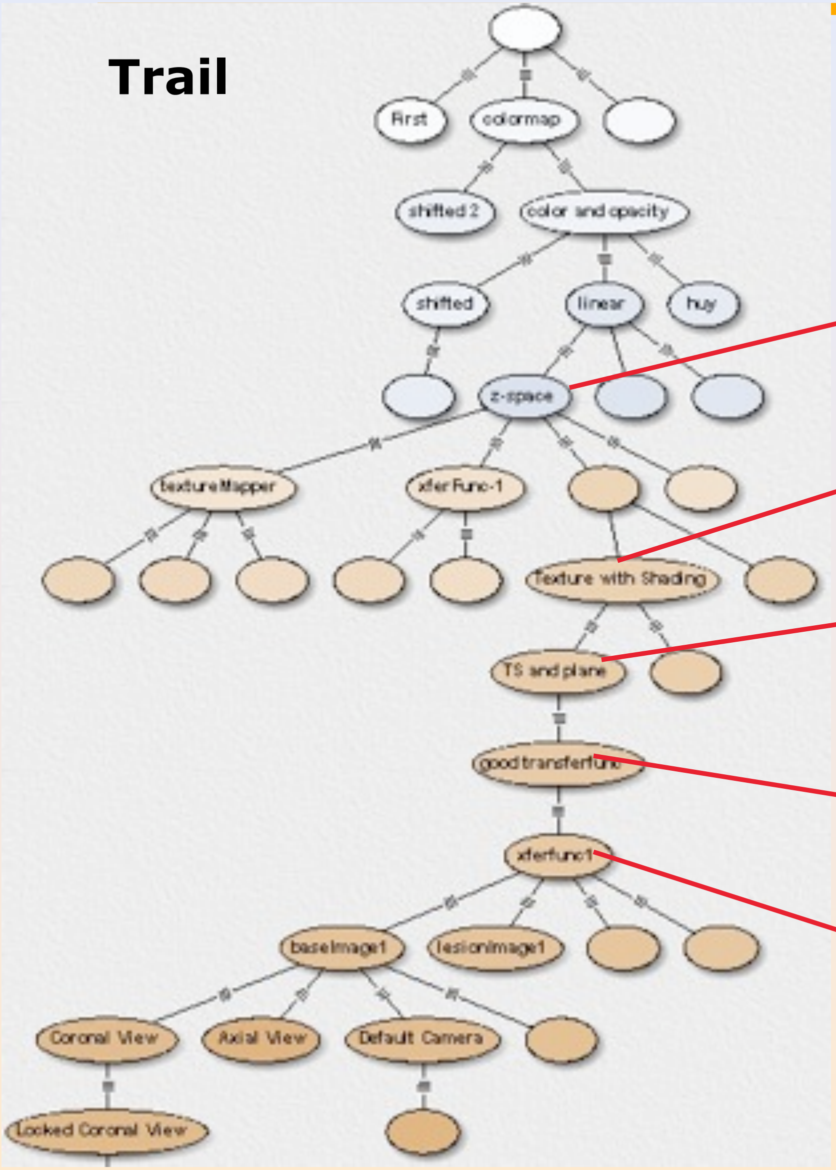
- Comprehensive *provenance infrastructure* for computational tasks
  - Data + workflow provenance
  - *Treat workflow as a 1st-class data product*
- Support for *exploratory* tasks such as visualization and data mining
  - Task specification iteratively refined as users generate and test hypotheses
- VisTrails manages the data, metadata and the exploration process, scientists can focus on *science!*
- Not a replacement for visualization or scientific workflow systems: infrastructure that can be combined with and enhance these systems
- Focus on usability—build tools for scientists

http://www.vistrails.org

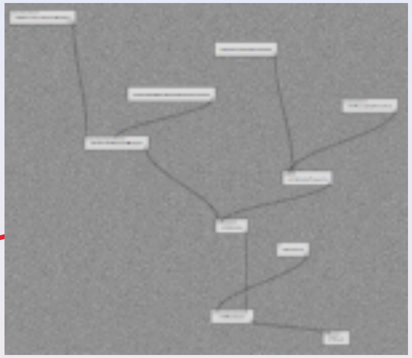
Keeping Exploration Trails



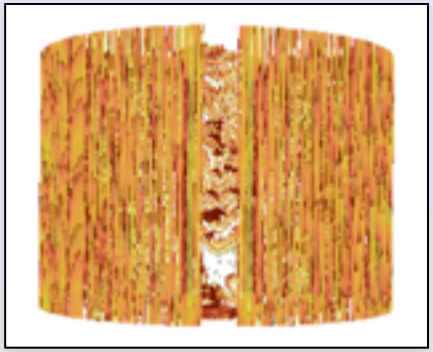
Keeping Exploration Trails



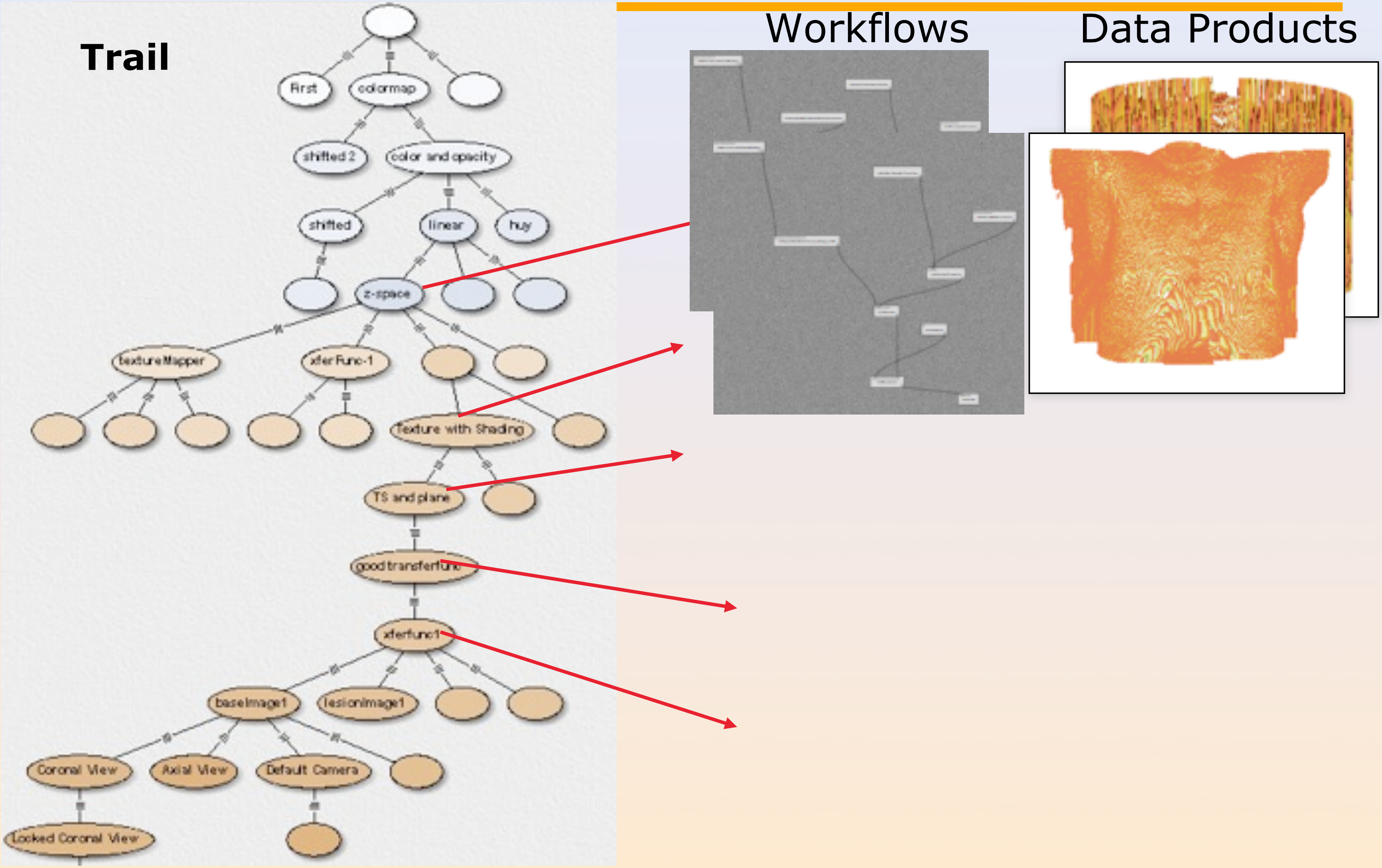
Workflows



Data Products



Keeping Exploration Trails

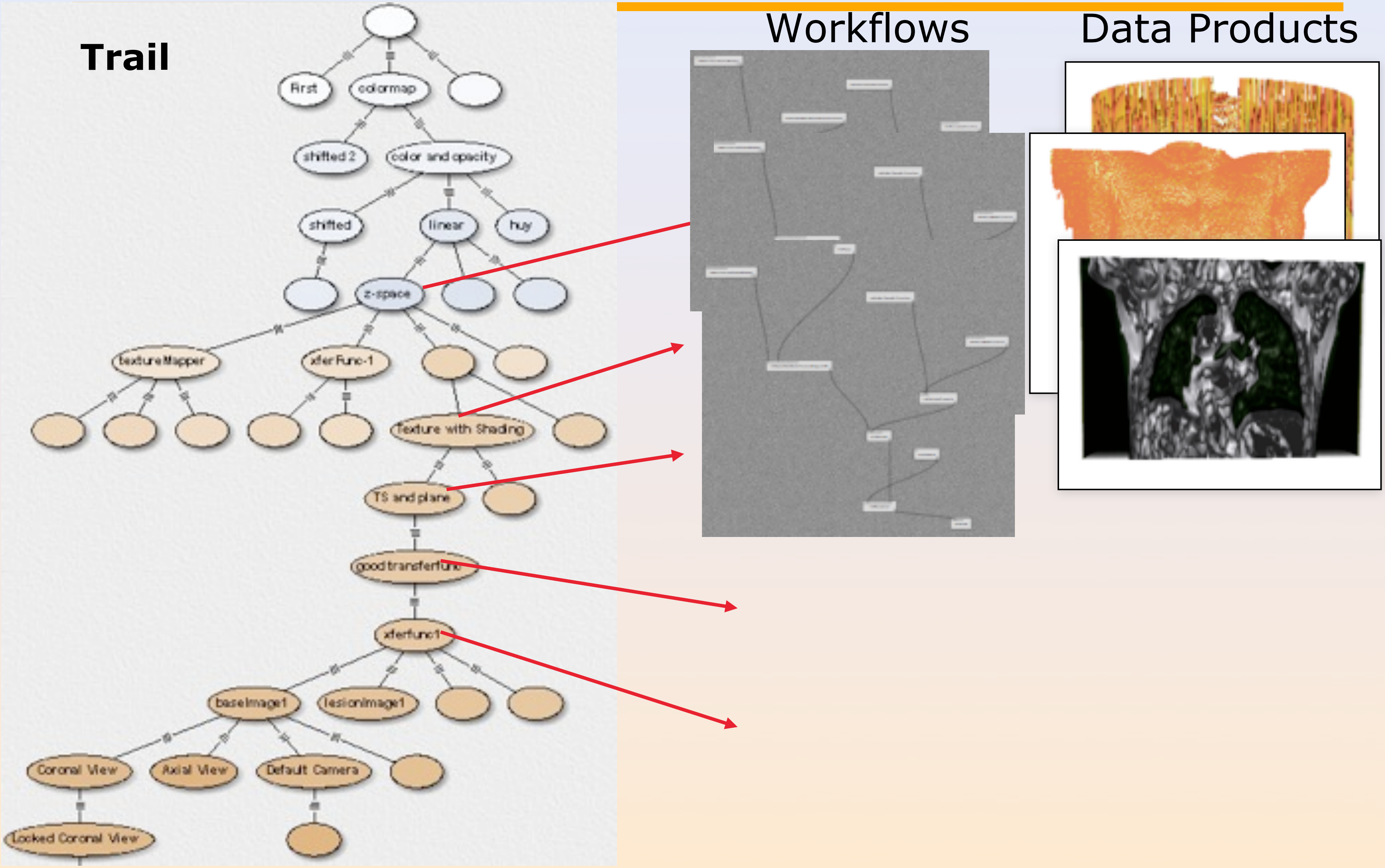


Trail

Workflows

Data Products

Keeping Exploration Trails

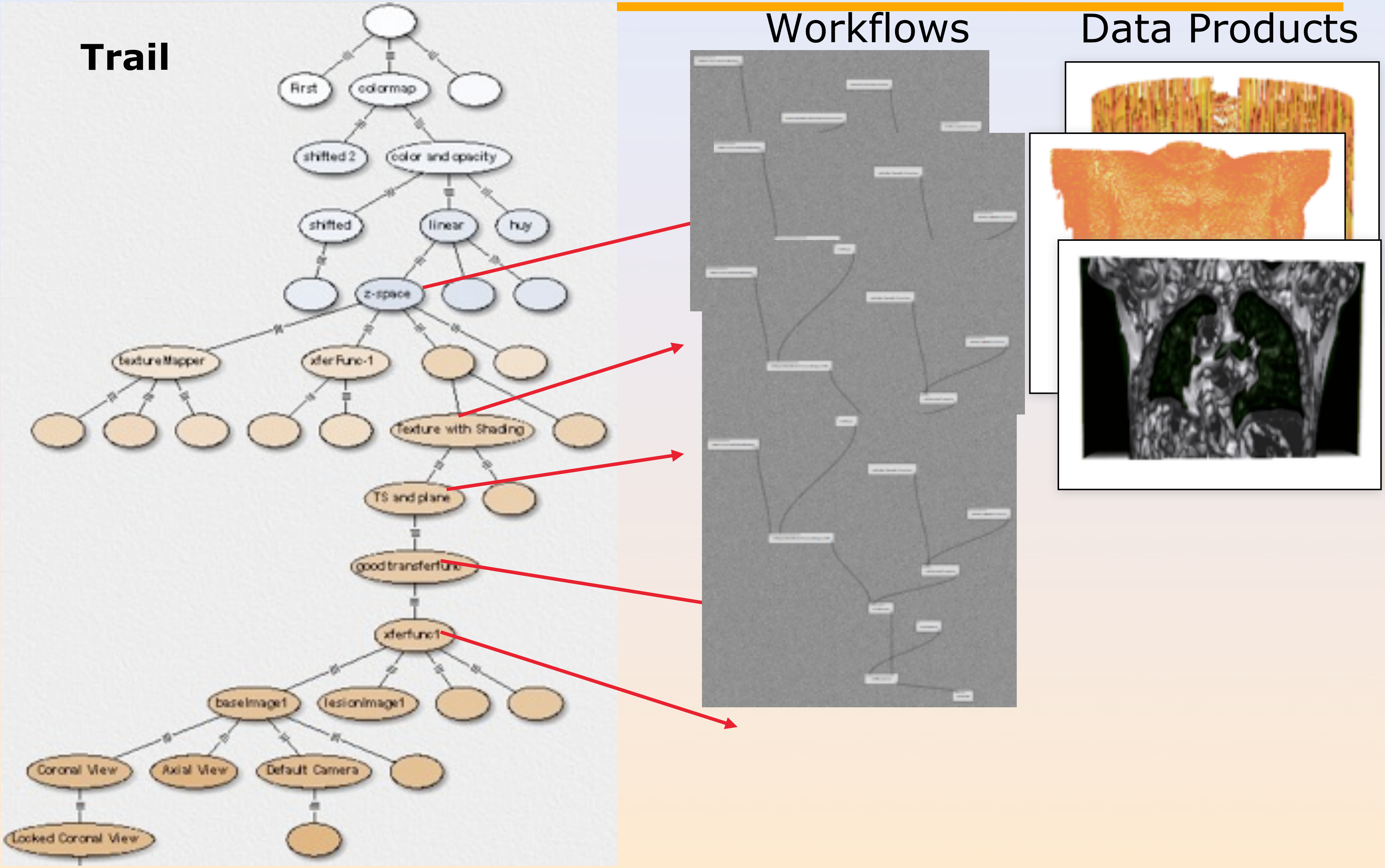


Trail

Workflows

Data Products

Keeping Exploration Trails

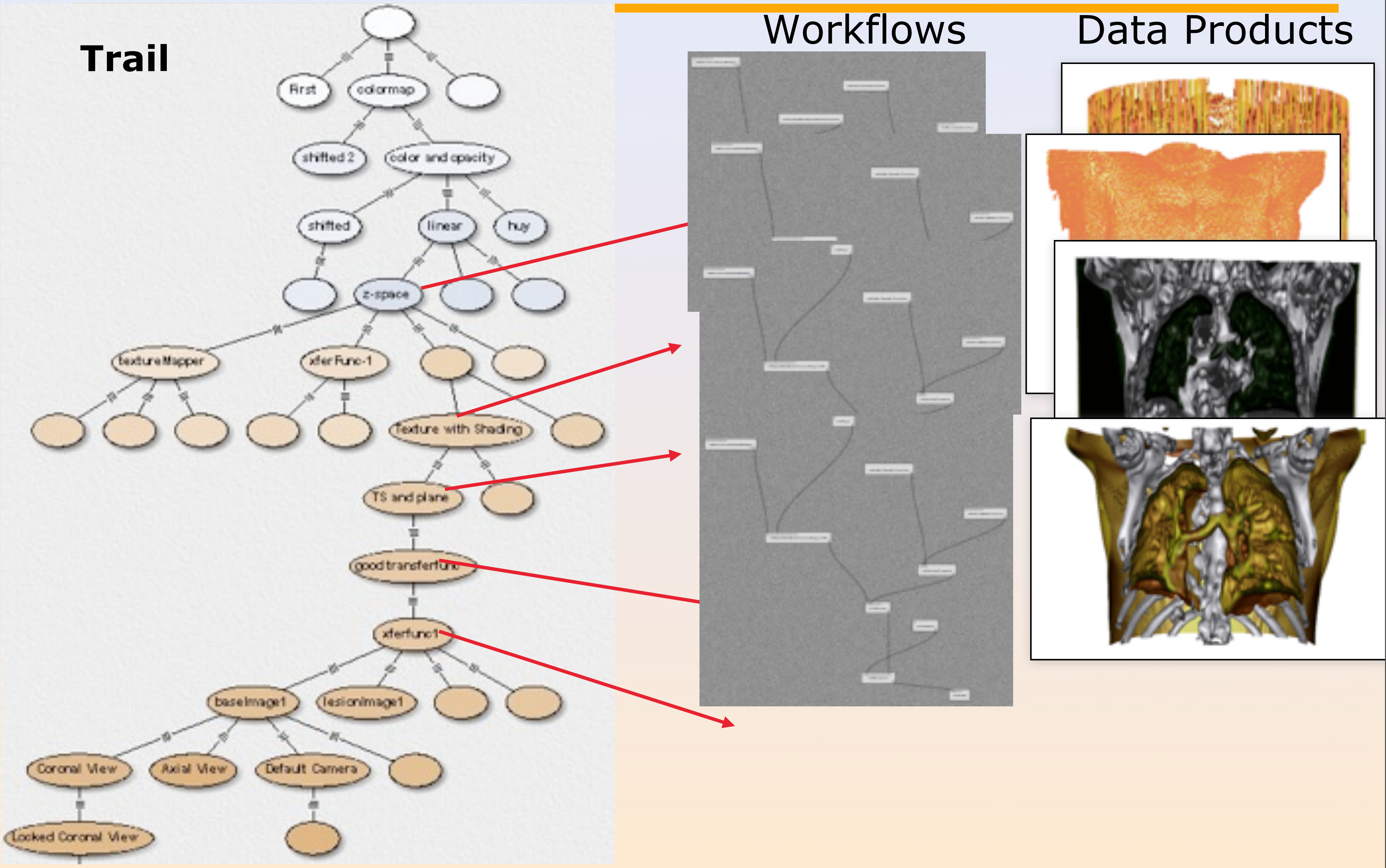


Trail

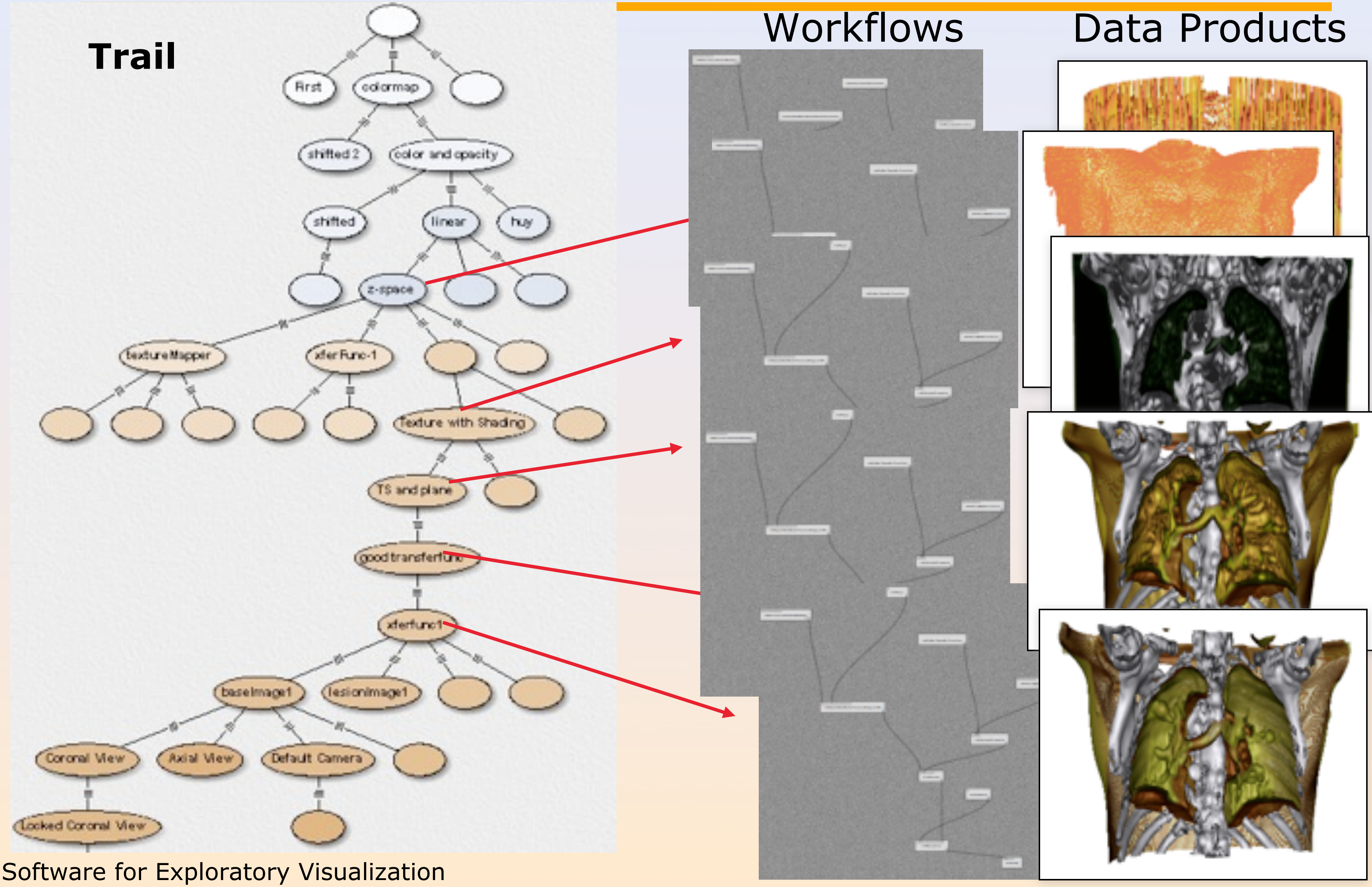
Workflows

Data Products

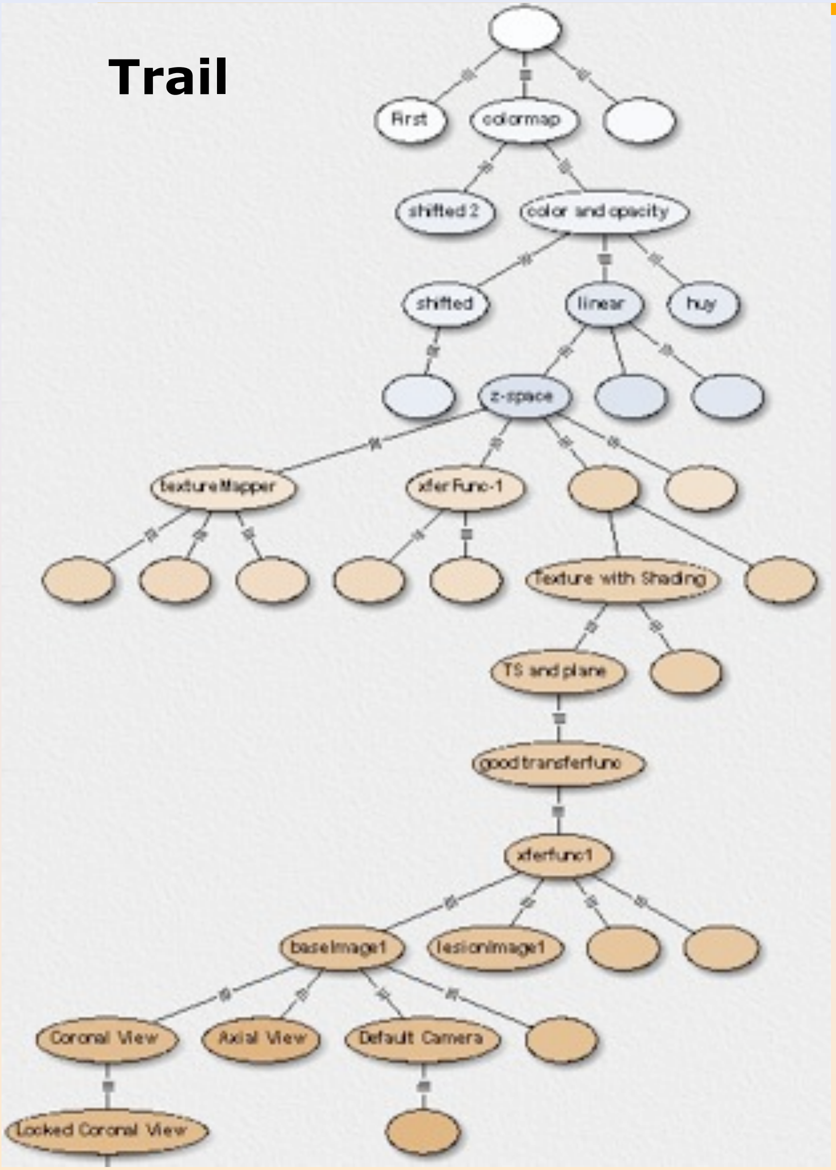
Keeping Exploration Trails



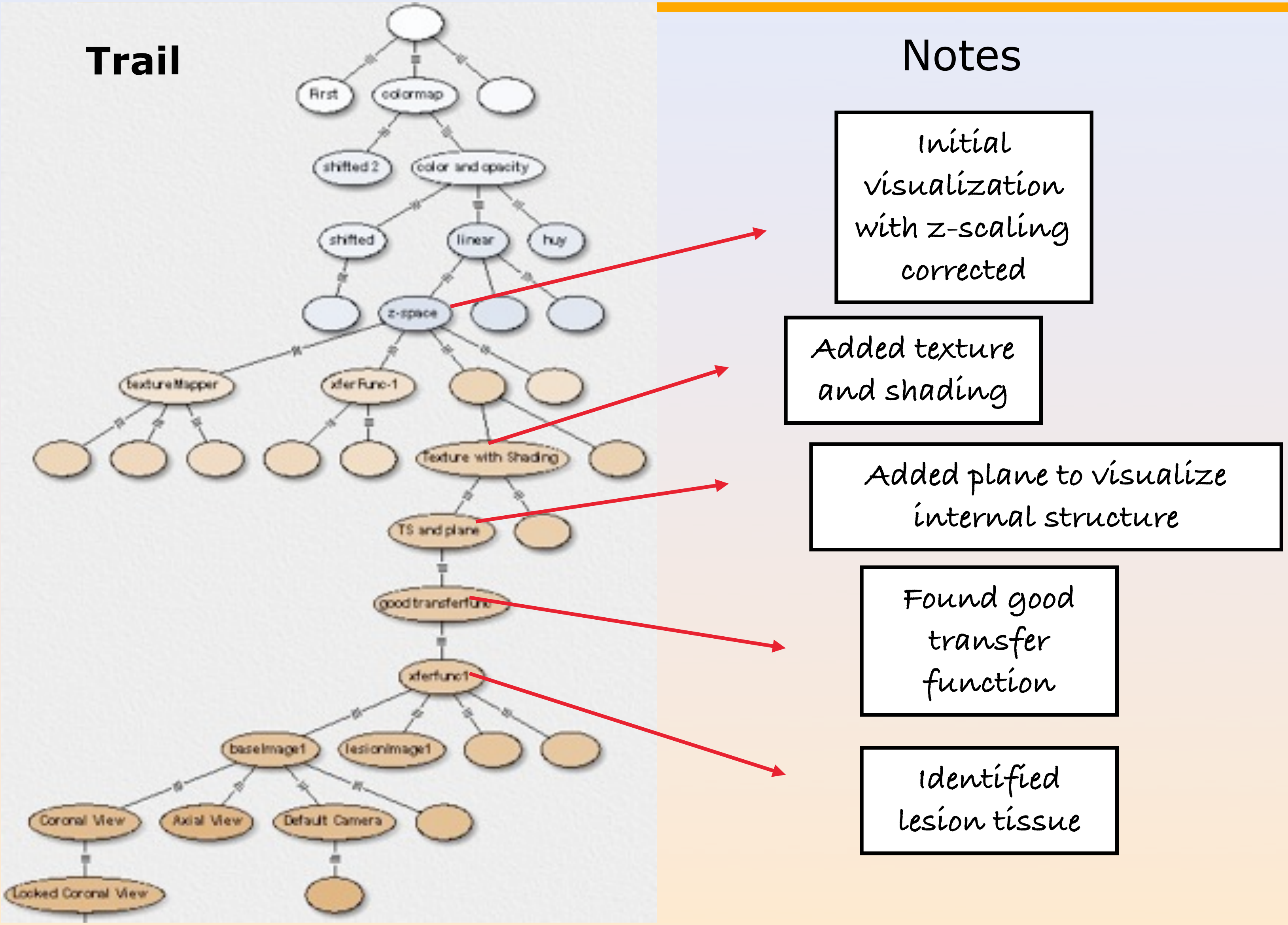
Keeping Exploration Trails



Keeping Exploration Trails



Keeping Exploration Trails



Trail

Notes

Initial visualization with z-scaling corrected

Added texture and shading

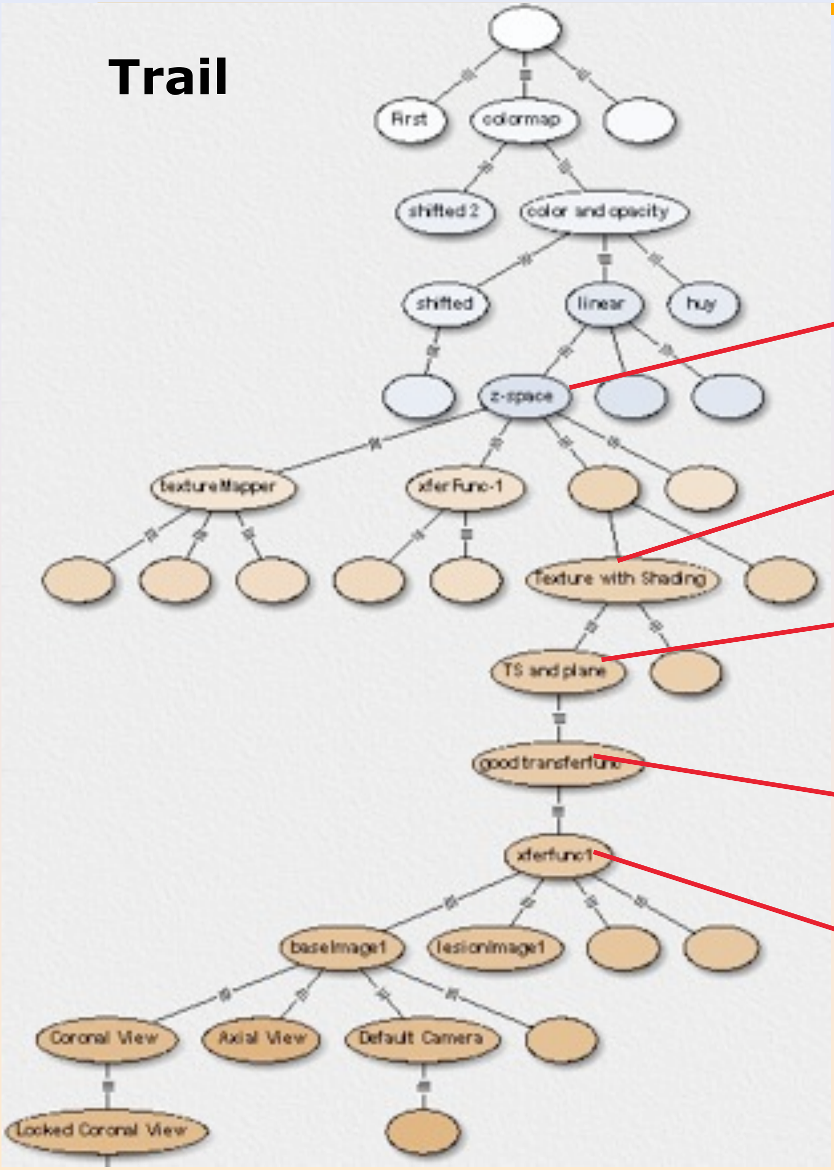
Added plane to visualize internal structure

Found good transfer function

Identified lesion tissue

Keeping Exploration Trails

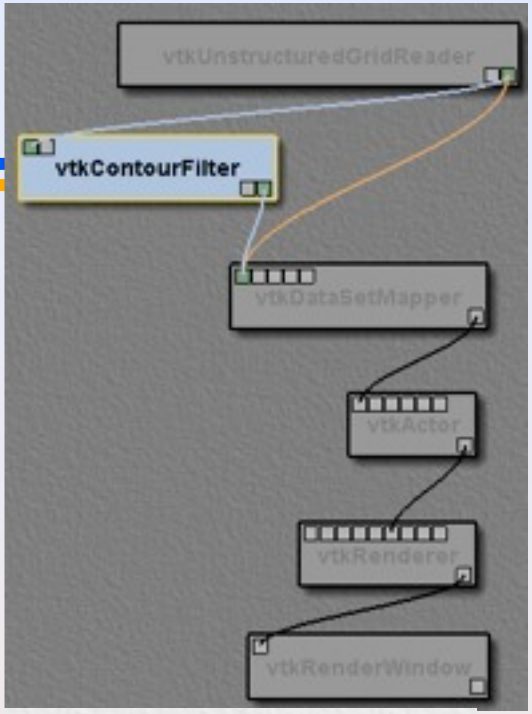
Trail



Notes	User
Initial visualization with z-scaling corrected	*juliana*
Added texture and shading	*eranders*
Added plane to visualize internal structure	*eranders*
Found good transfer function	*eranders*
Identified lesion tissue	*stevec*

Change-Based Provenance

- Records actions
- Provenance = changes to computational tasks
  - Add a module, add a connection, change a parameter value
- Extensible *change* algebra



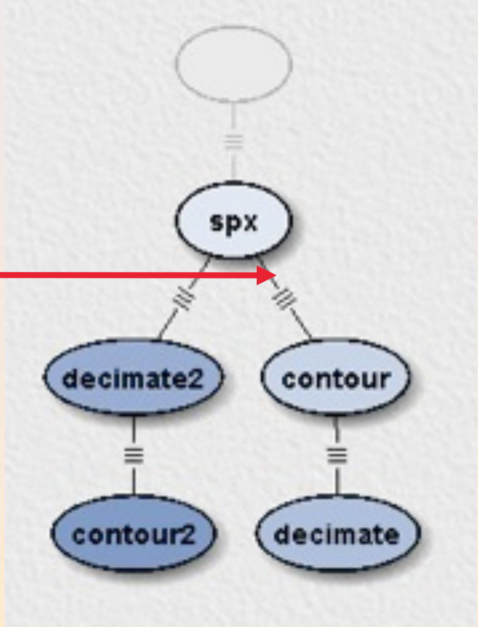
addModule

deleteConnection

addConnection

addConnection

setParameter

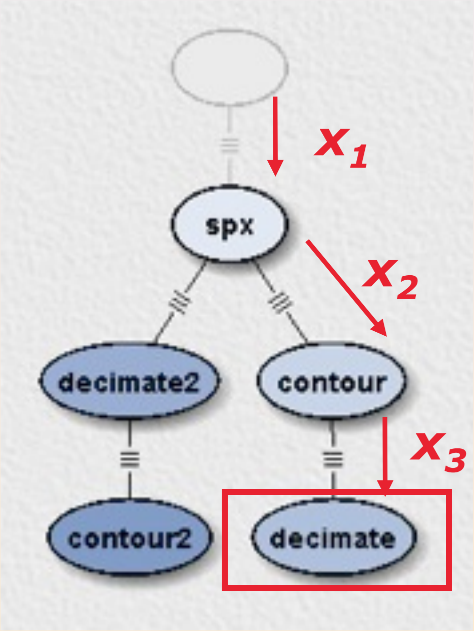


Change-Based Provenance

- Records actions
- Provenance = changes to computational tasks
  - Add a module, add a connection, change a parameter value
- Extensible *change* algebra
- A *vistrail* node v_t corresponds to the workflow that is constructed by the sequence of actions from the root to v_t

$$v_t = x_n \circ x_{n-1} \circ \ldots \circ x_1 \circ \emptyset$$

vistrail



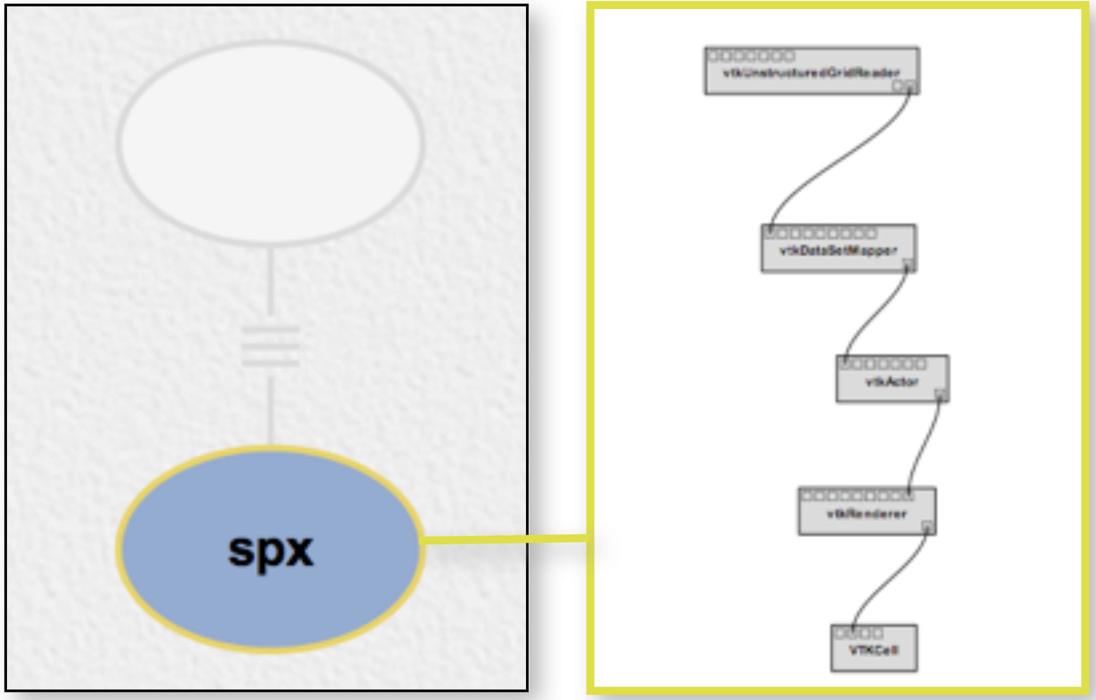
[Freire et al, IPAW 2006]

Change-based provenance

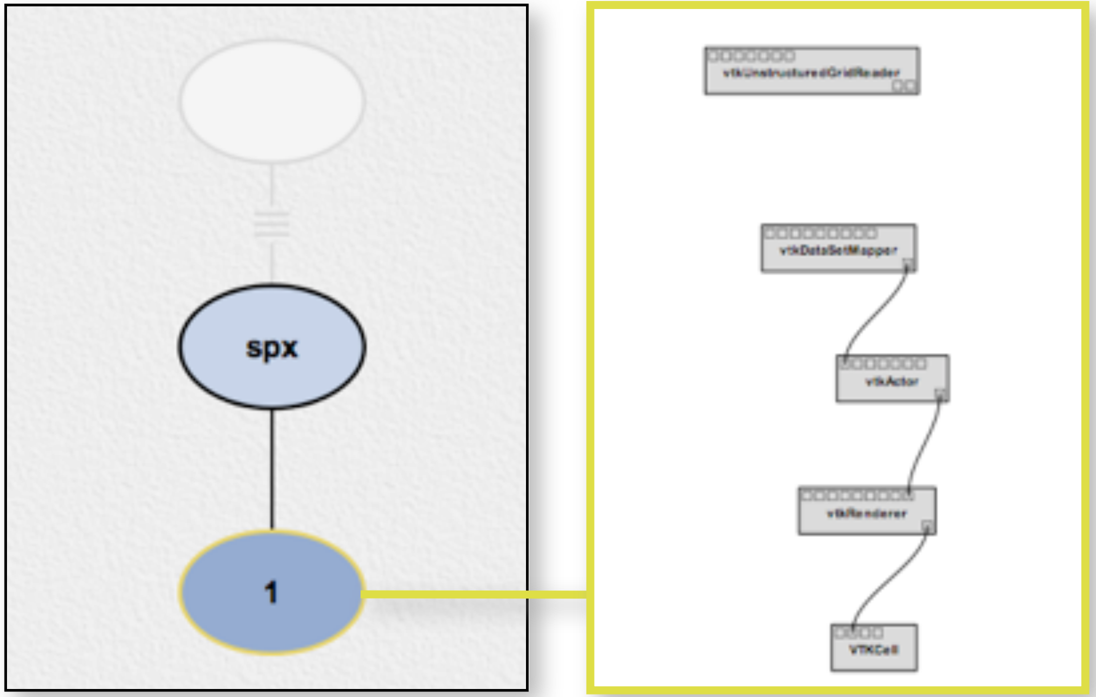
- Records changes to workflows
- Workflow evolution is captured in a vistrail, a rooted tree where

–nodes correspond to workflow versions

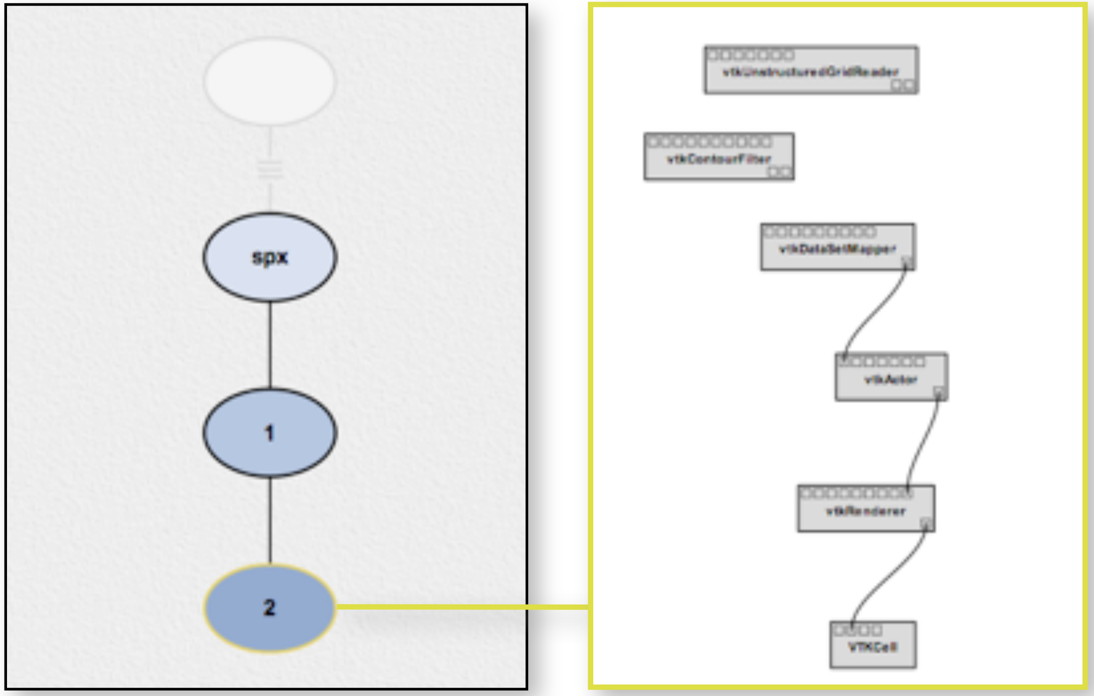
–edges correspond to actions that transform the parent into the child workflow



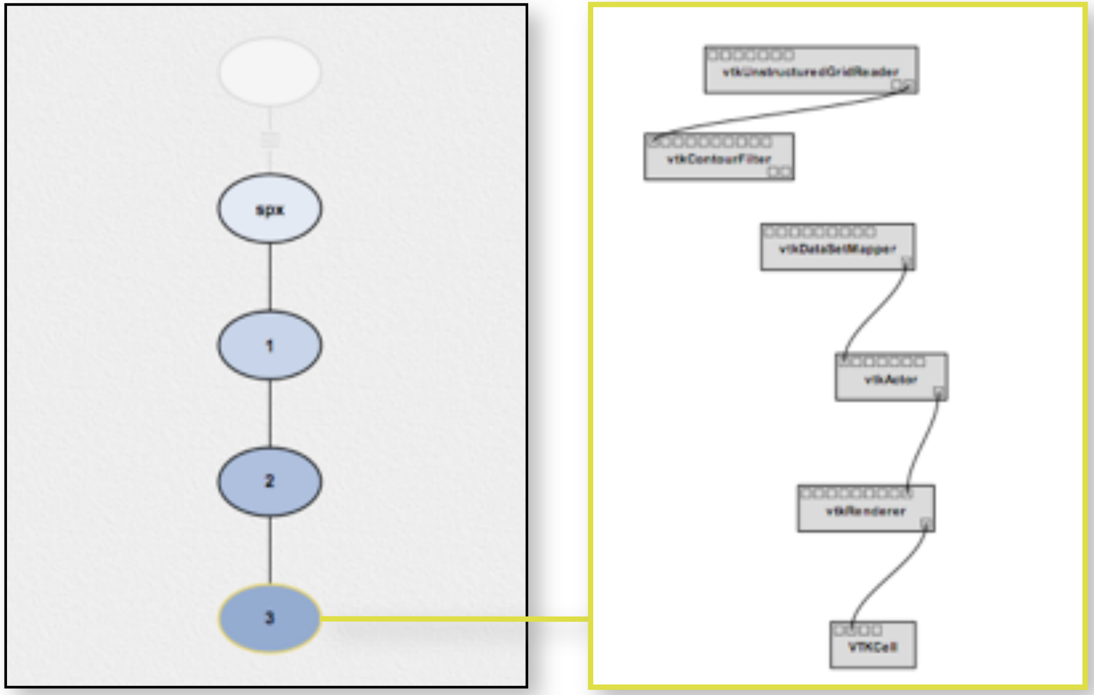
Change-based provenance



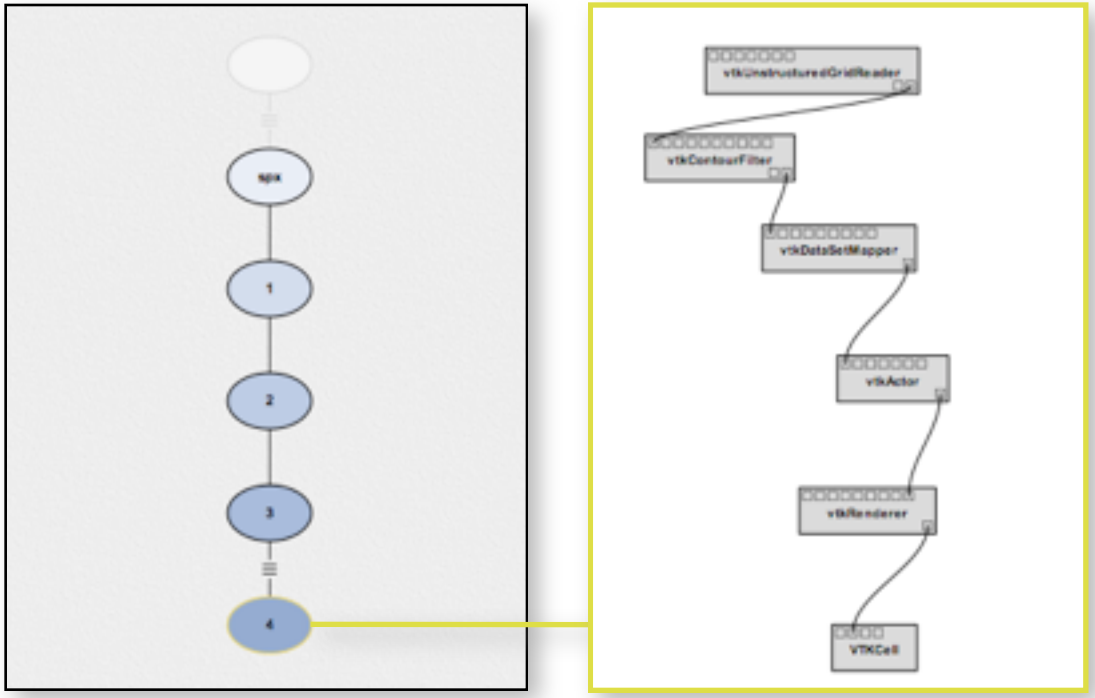
Change-based provenance



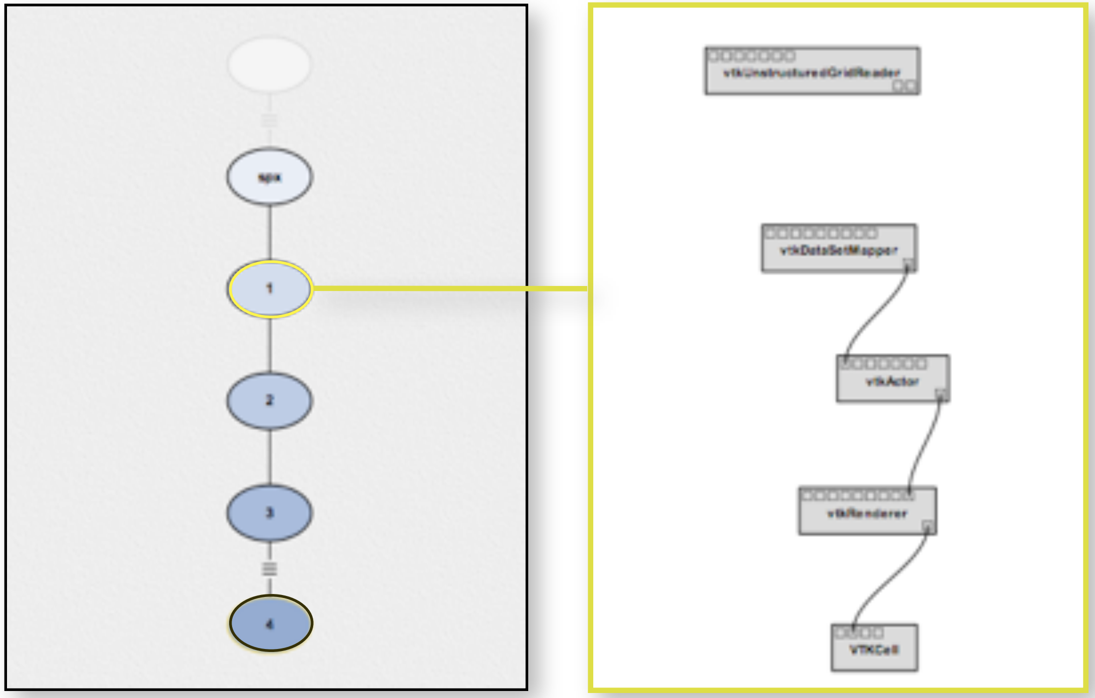
Change-based provenance



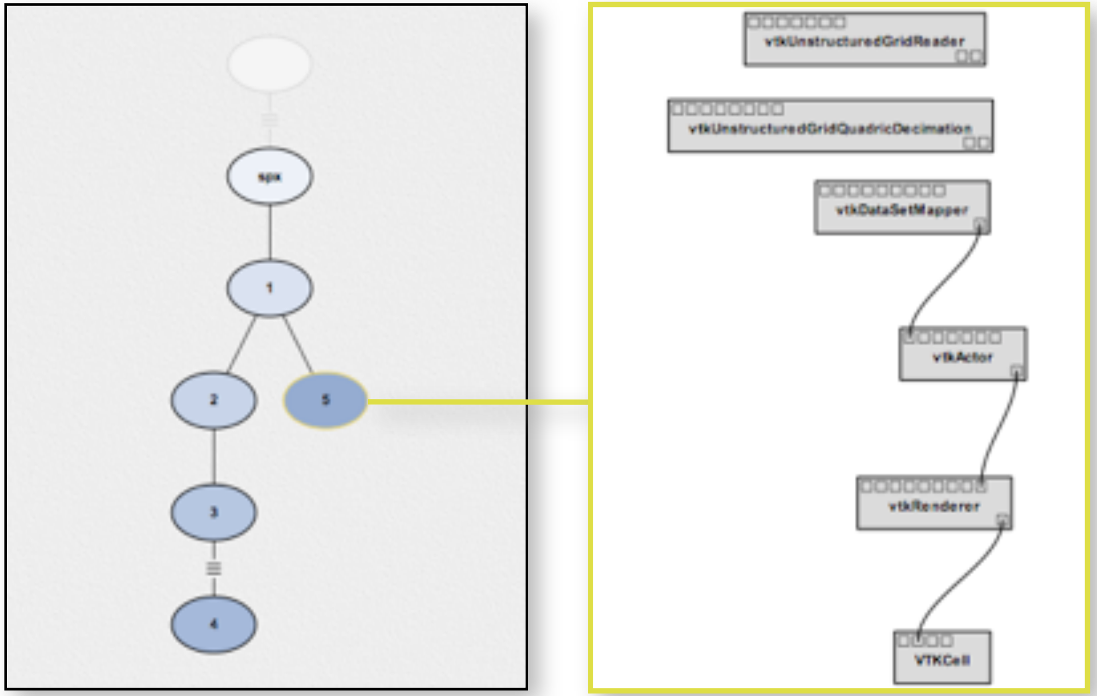
Change-based provenance



Change-based provenance



Change-based provenance

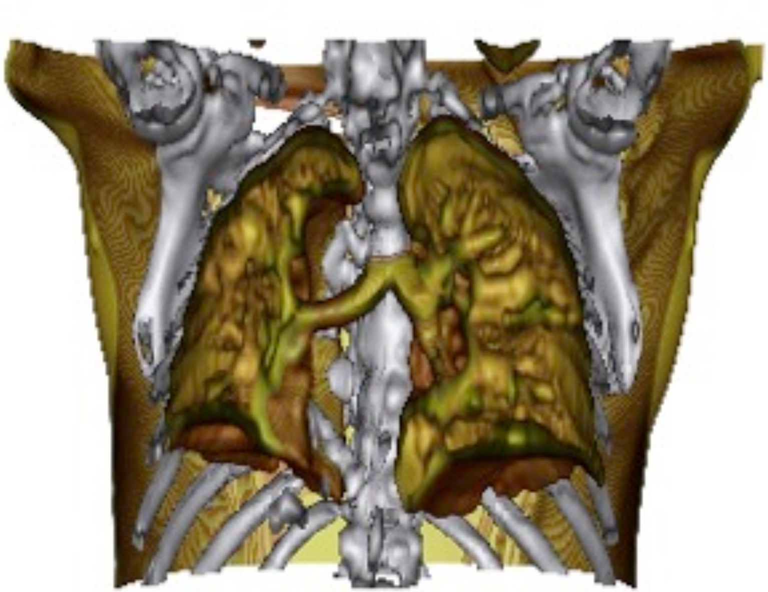

spx
1
2
5
3
4
vtkUnstructuredGridReader
vtkUnstructuredGridQuadricDecimation
vtkDataSetMapper
vtkActor
vtkRenderer
VTKCell

Change-Based Provenance

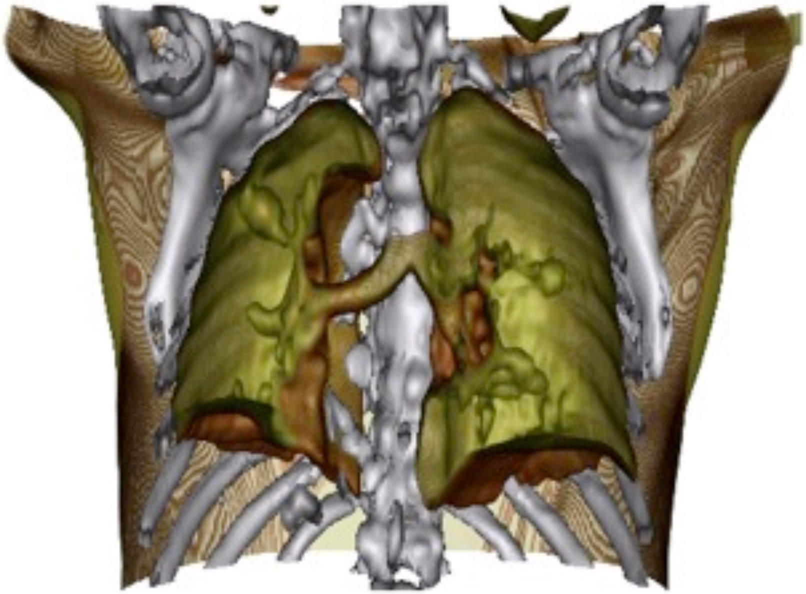
- General: Works with any system that has undo/redo!
- Concise representation
- Uniformly captures data and workflow provenance
  - Data provenance: where does a specific data product come from?
  - Workflow evolution: how has workflow structure changed over time?
- Detailed information about the exploration process
  - Results can be reproduced
- Provenance beyond reproducibility:
  - Scientists can return to any point in the exploration space
  - Enables scalable exploration of the parameter space

  (and compare results using a spreadsheet)
  - Support for collaboration
  - Understand problem-solving strategies—knowlegde re-use

What’s the difference?

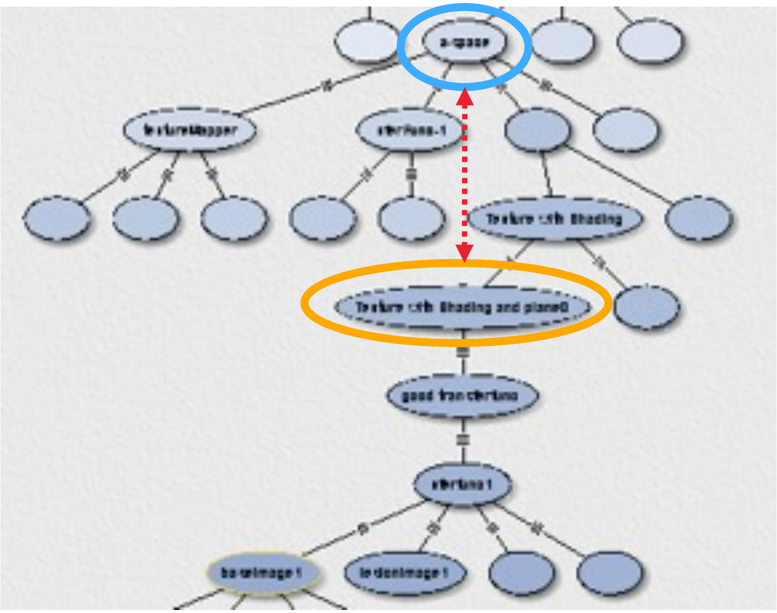


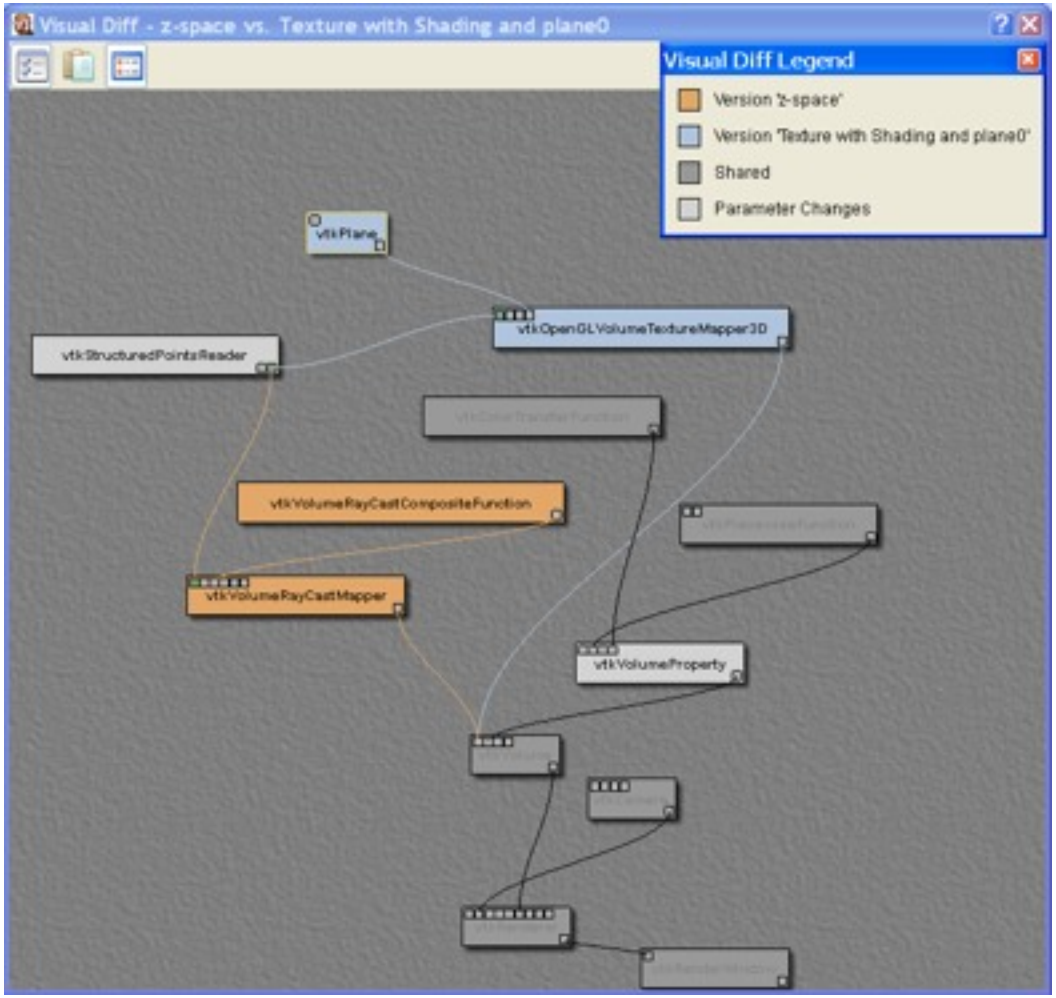
baseImage1



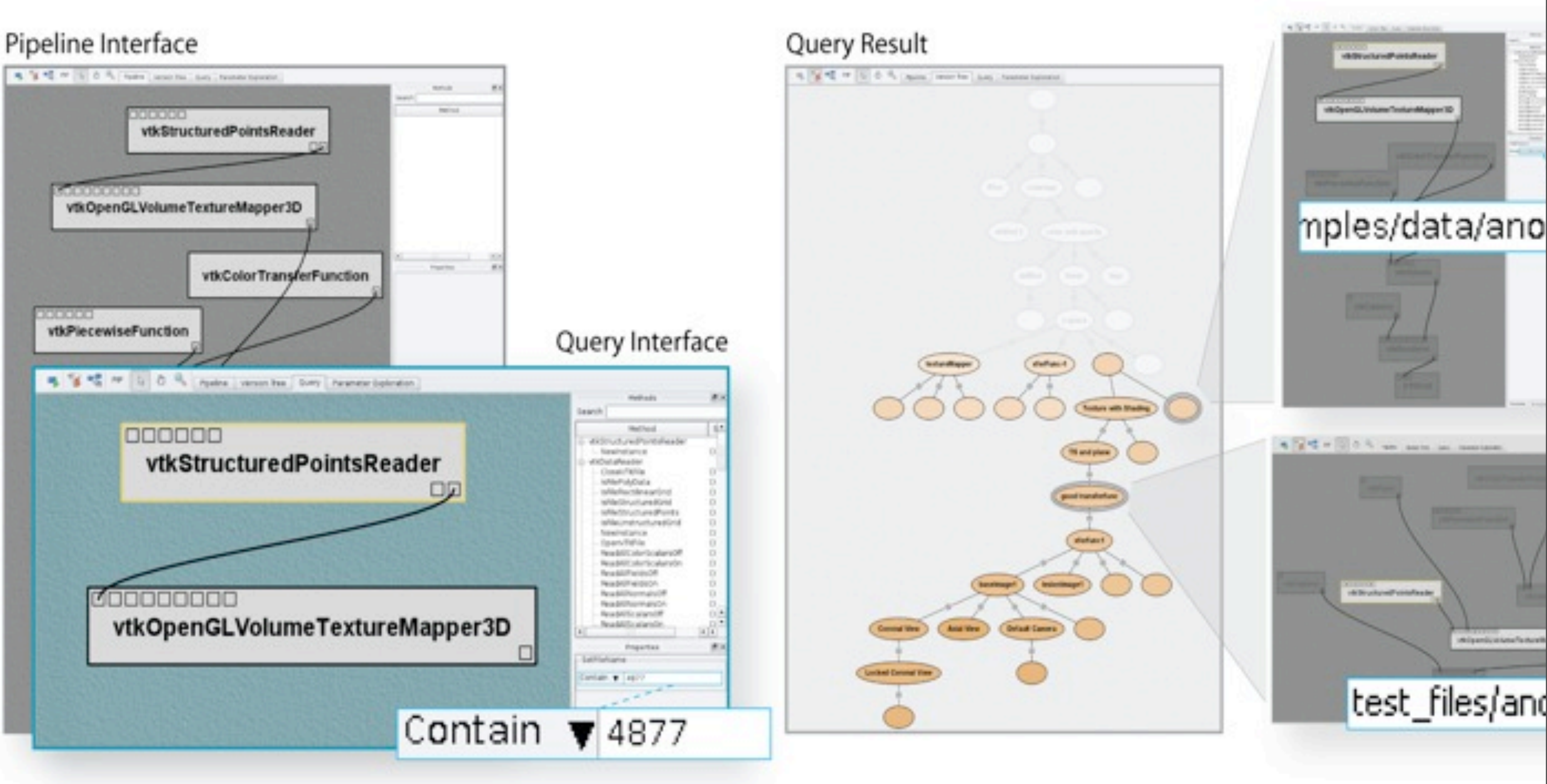
lesionImage1

Differences in Specification

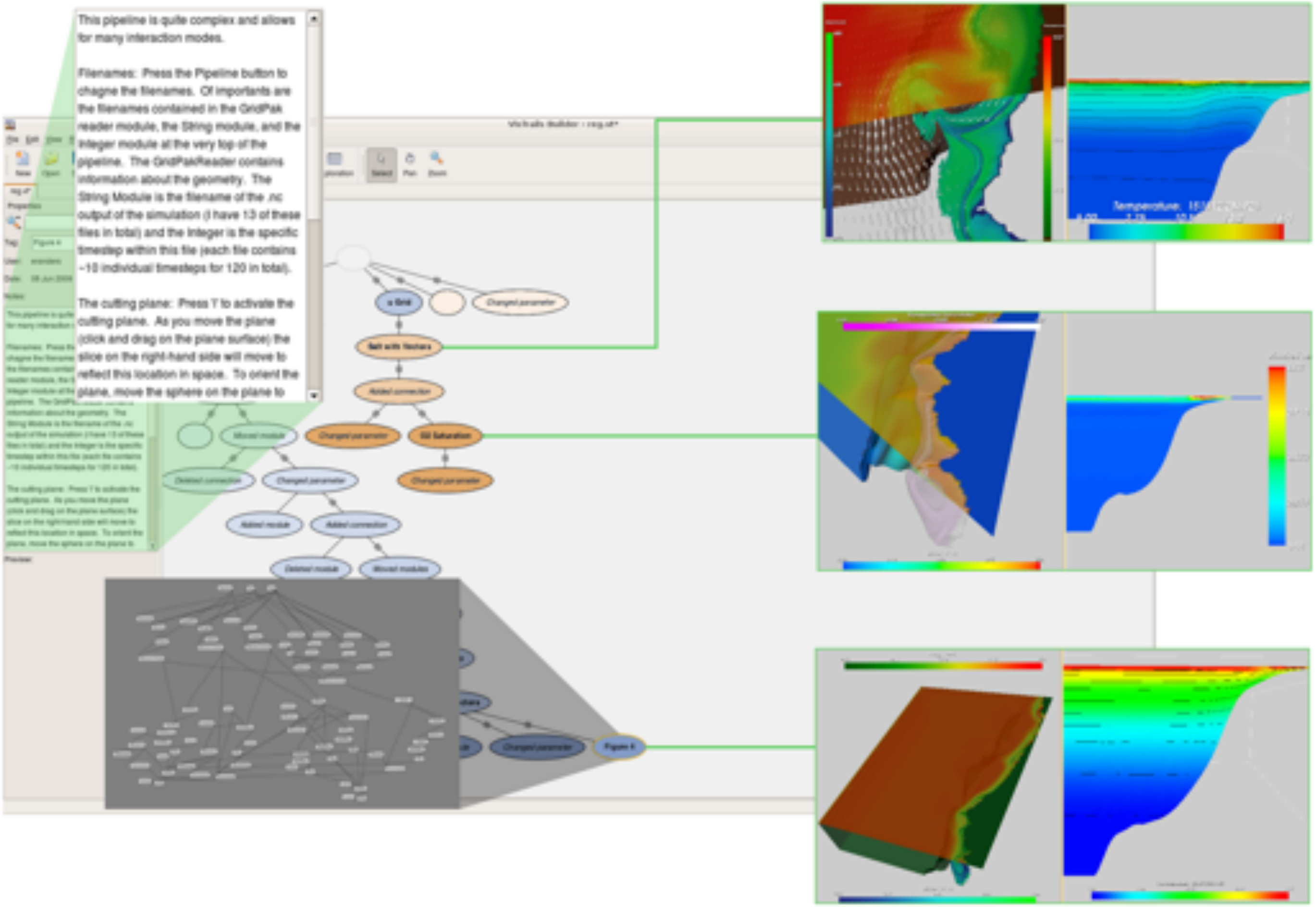




Query by Example



VisTrails: Provenance of Exploration



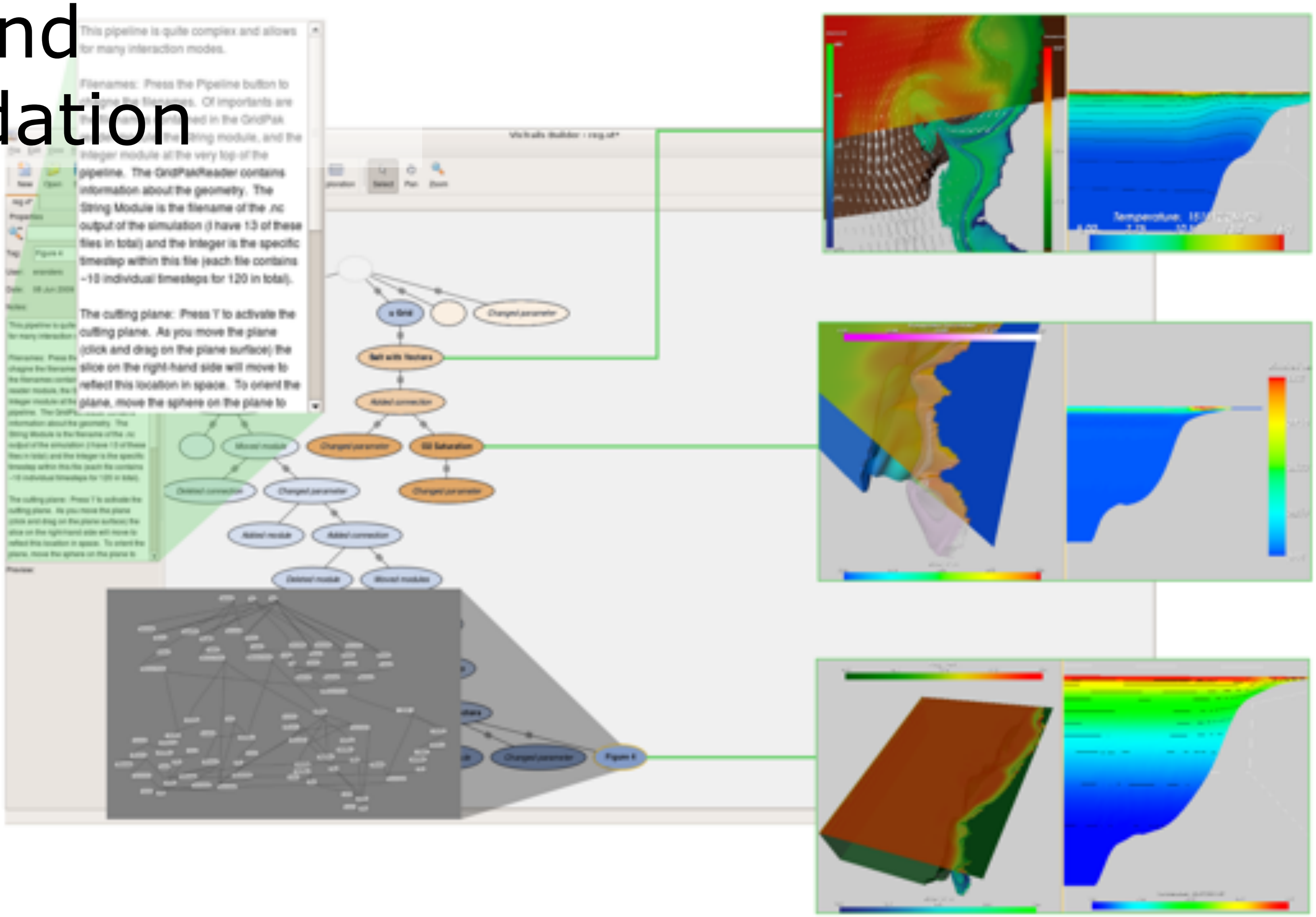
This pipeline is quite complex and allows for many interaction modes.

Filenames: Press the Pipeline button to chagne the filenames. Of importants are the filenames contained in the GridPak reader module, the String module, and the integer module at the very top of the pipeline. The GridPakReader contains information about the geometry. The String Module is the filename of the .nc output of the simulation (I have 13 of these files in total) and the integer is the specific timestep within this file (each file contains ~10 individual timesteps for 120 in total).

The cutting plane: Press 'i' to activate the cutting plane. As you move the plane (click and drag on the plane surface) the slice on the right-hand side will move to reflect this location in space. To orient the plane, move the sphere on the plane to

VisTrails: Provenance of Exploration

Reproducibility
and
Validation



This pipeline is quite complex and allows for many interaction modes.

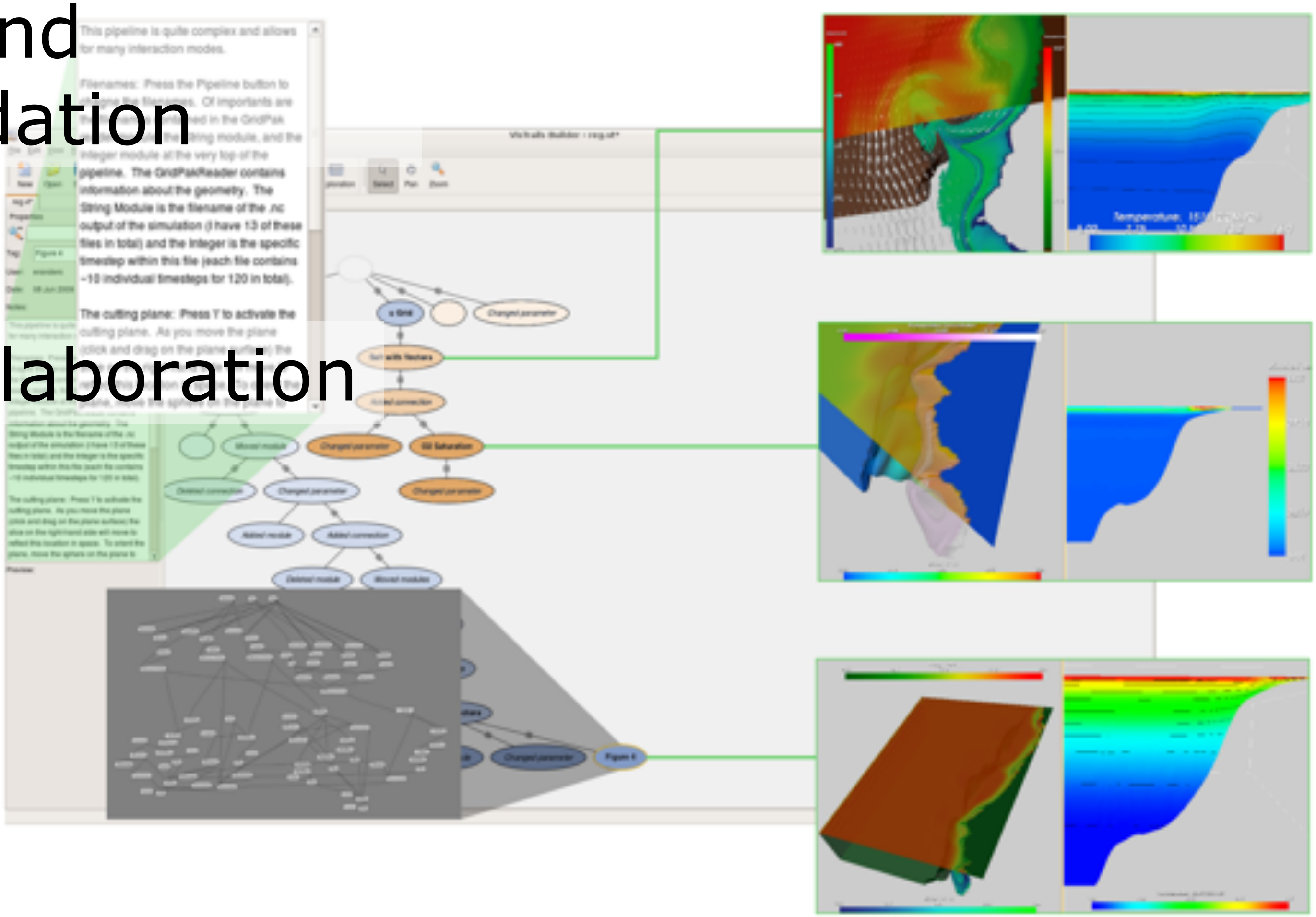
Filenames: Press the Pipeline button to change the filenames. Of importants are the filenames contained in the GridPak reader module, the String module, and the Integer module at the very top of the pipeline. The GridPakReader contains information about the geometry. The String Module is the filename of the .nc output of the simulation (I have 13 of these files in total) and the Integer is the specific timestep within this file (each file contains ~10 individual timesteps for 120 in total).

The cutting plane: Press 'I' to activate the cutting plane. As you move the plane (click and drag on the plane surface) the slice on the right-hand side will move to reflect this location in space. To orient the plane, move the sphere on the plane to

VisTrails: Provenance of Exploration

Reproducibility
and
Validation

Collaboration



This pipeline is quite complex and allows for many interaction modes.

Filenames: Press the Pipeline button to change the filenames. Of importants are the filenames contained in the GridPak [illegible] the String module, and the integer module at the very top of the pipeline. The GridPakReader contains information about the geometry. The String Module is the filename of the .nc output of the simulation (I have 13 of these files in total) and the integer is the specific timestep within this file (each file contains ~10 individual timesteps for 120 in total).

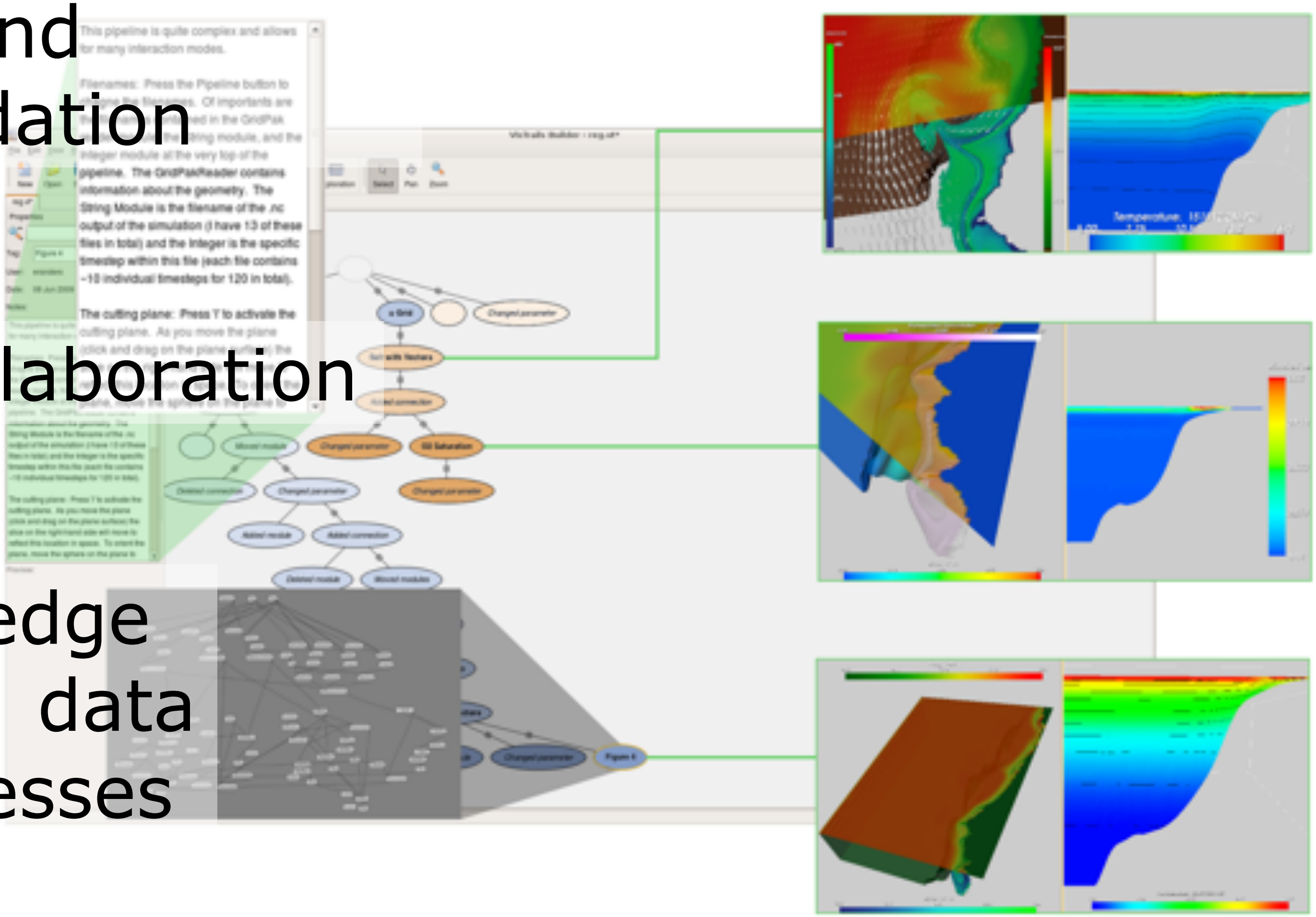
The cutting plane: Press 'i' to activate the cutting plane. As you move the plane (click and drag on the plane surface) the [illegible] reflect this location in space. To orient the plane, move the sphere on the plane to

VisTrails: Provenance of Exploration

Reproducibility and Validation

Collaboration

Knowledge sharing: data + processes



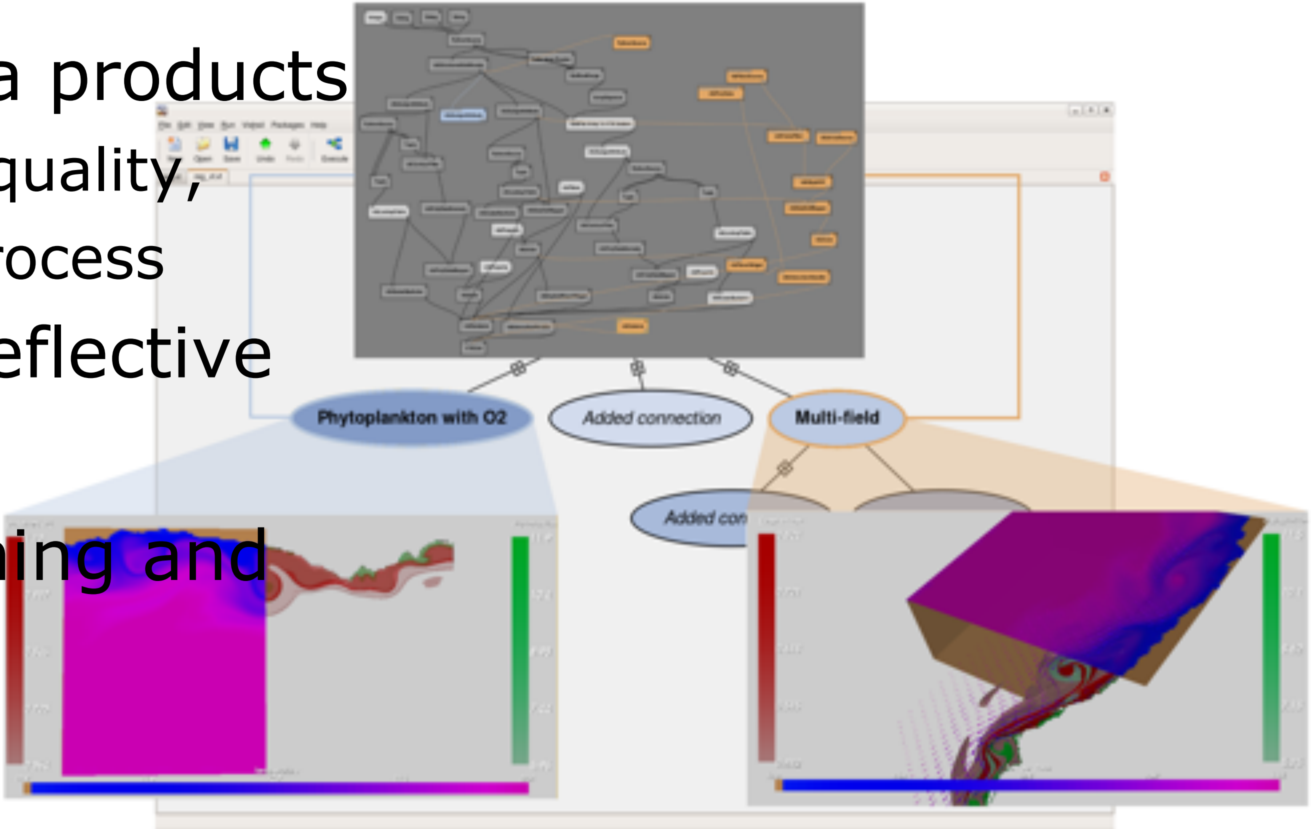
This pipeline is quite complex and allows for many interaction modes.

Filenames: Press the Pipeline button to change the filenames. Of importants are the filenames contained in the GridPak reader module, the string module, and the integer module at the very top of the pipeline. The GridPakReader contains information about the geometry. The String Module is the filename of the .nc output of the simulation (I have 13 of these files in total) and the integer is the specific timestep within this file (each file contains ~10 individual timesteps for 120 in total).

The cutting plane: Press 'i' to activate the cutting plane. As you move the plane (click and drag on the plane surface) the

Benefits of Provenance

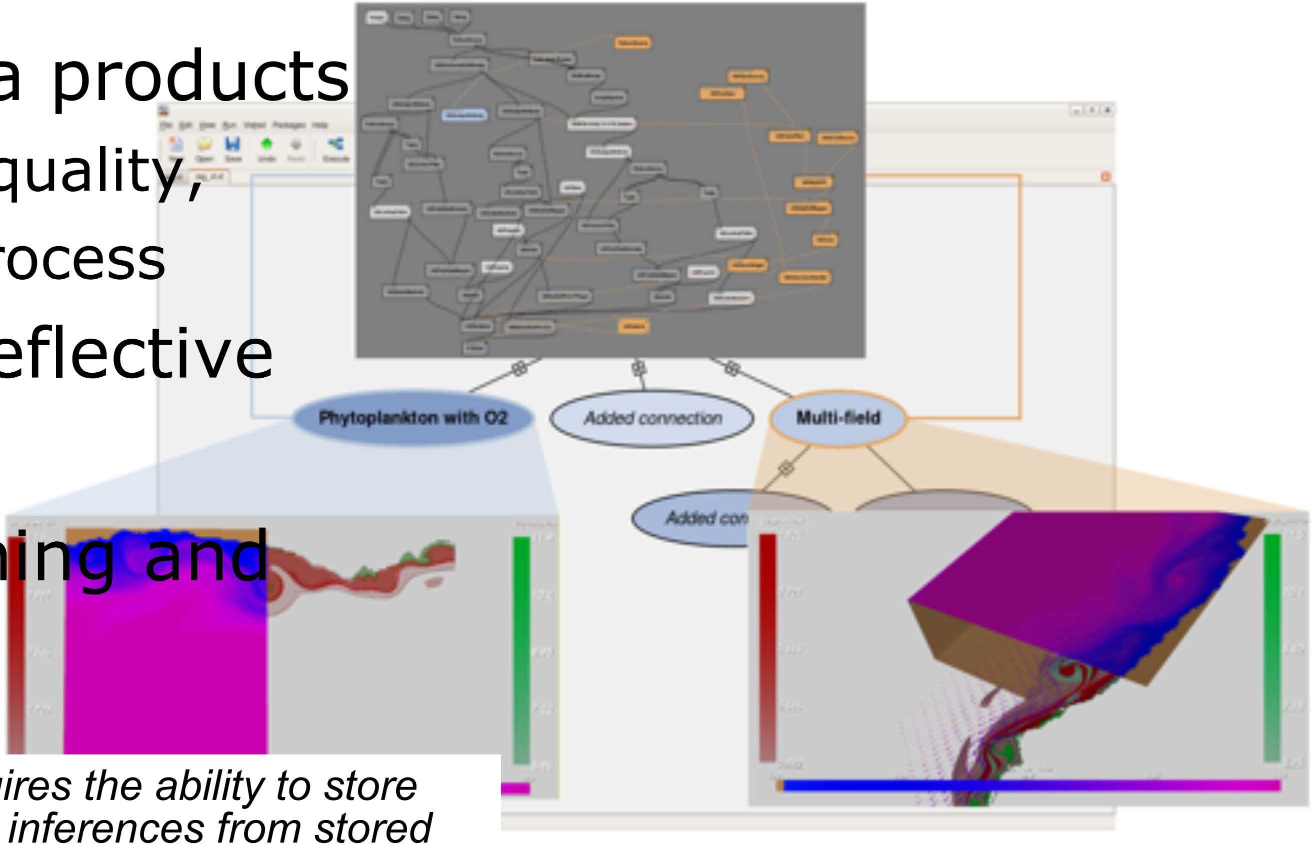
- Compare data products
  - Assess data quality, understand process
- Support for reflective reasoning
- Improve training and teaching



[Freire et al., IPAW 2006]

Benefits of Provenance

- Compare data products
  - Assess data quality, understand process
- Support for reflective reasoning
- Improve training and teaching

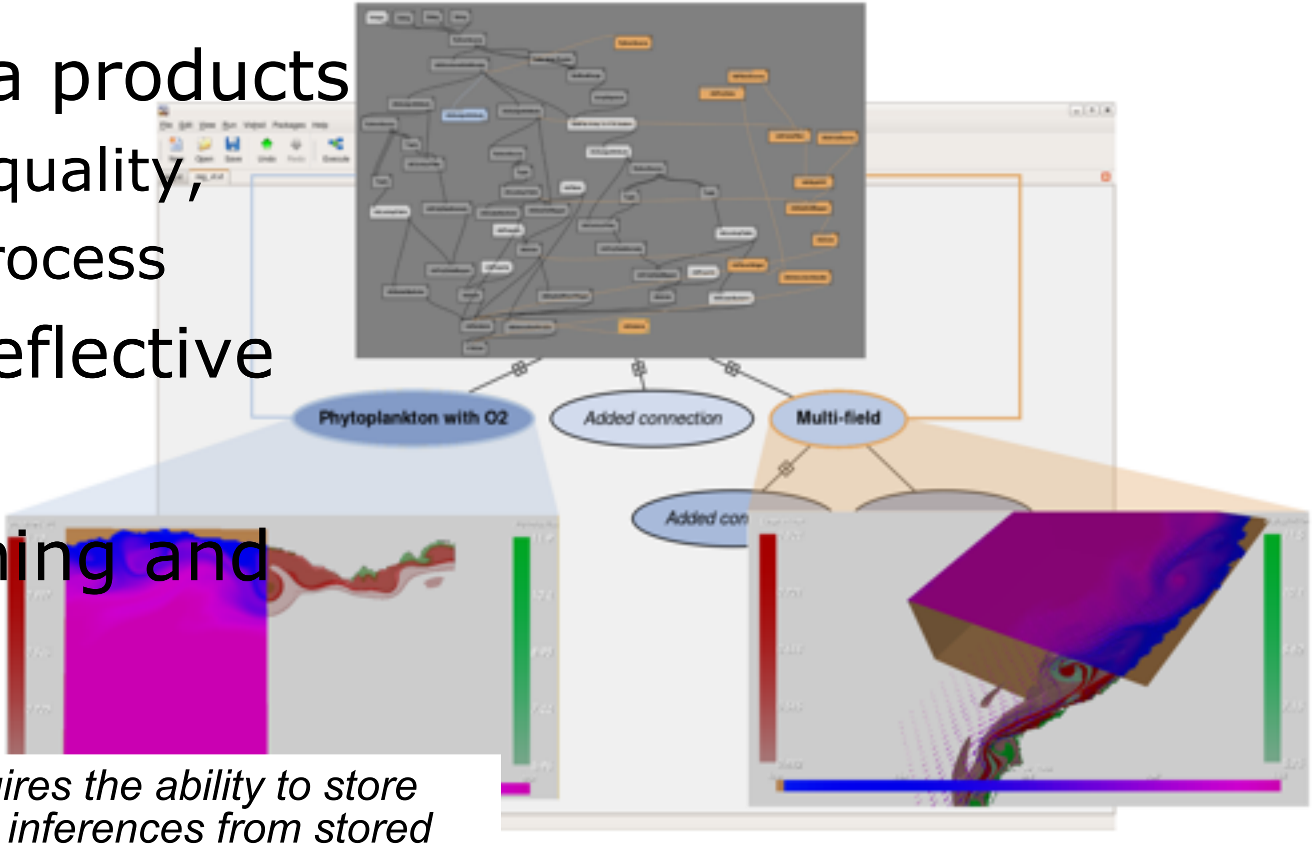


"Reflective reasoning requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. …the process is slow and laborious"
Things that make us smart, Donald A. Norman

[Freire et al., IPAW 2006]

Benefits of Provenance

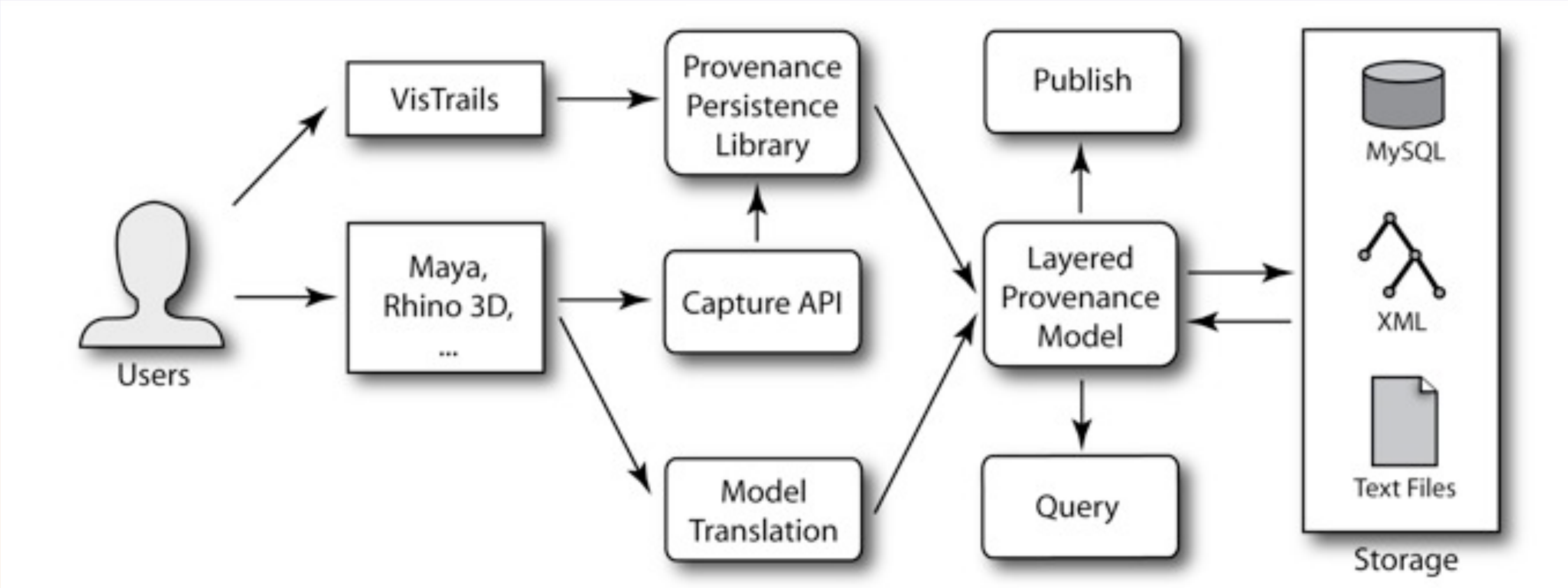
- Compare data products
  - Assess data quality, understand process
- Support for reflective reasoning
- Improve training and teaching



"Reflective reasoning requires the ability to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward, sometimes backtracking when a promising line of thought proves to be unfruitful. …the process is slow and laborious"
Things that make us smart, Donald A. Norman

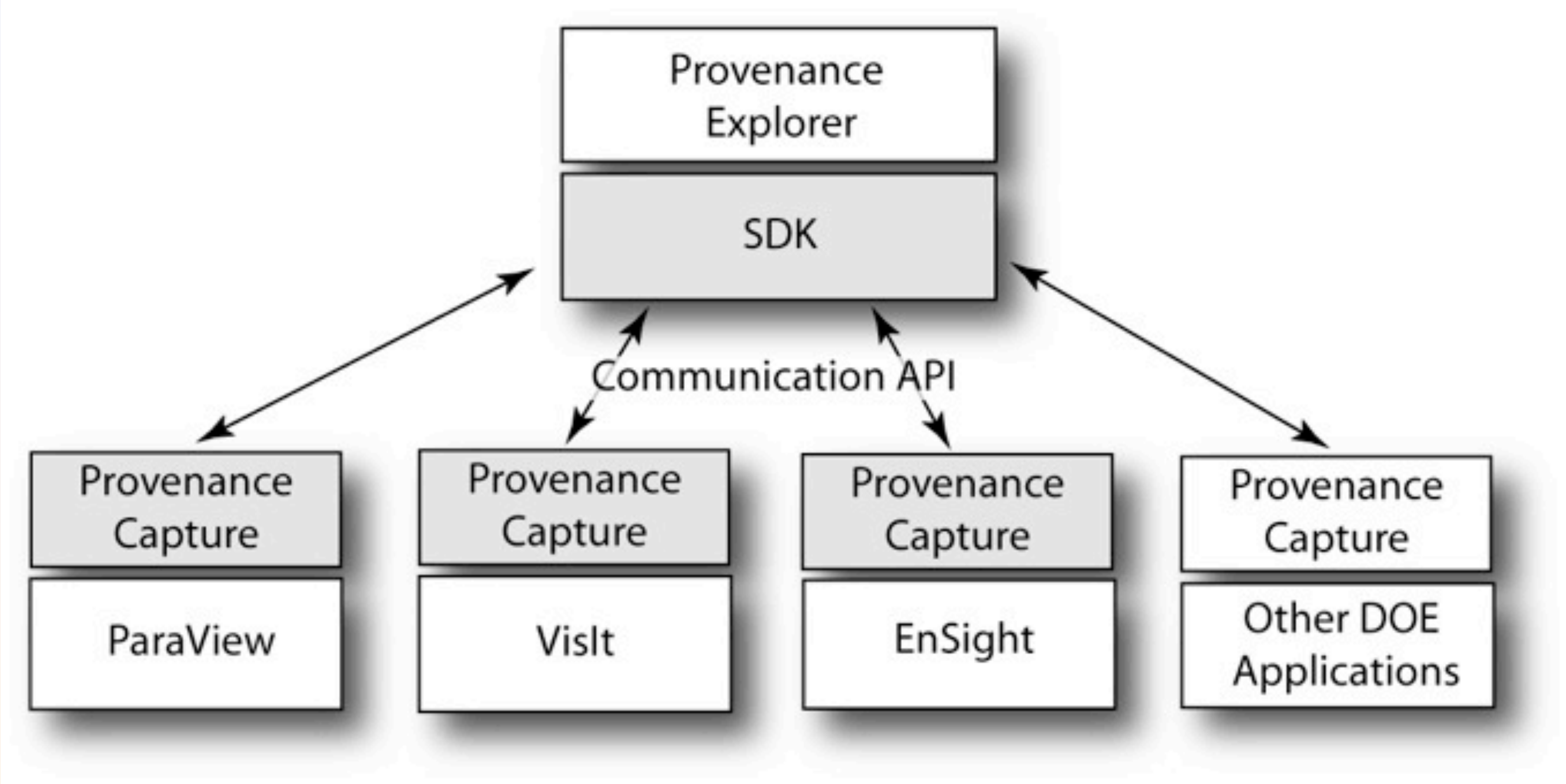
[Freire et al., IPAW 2006]

Provenance API

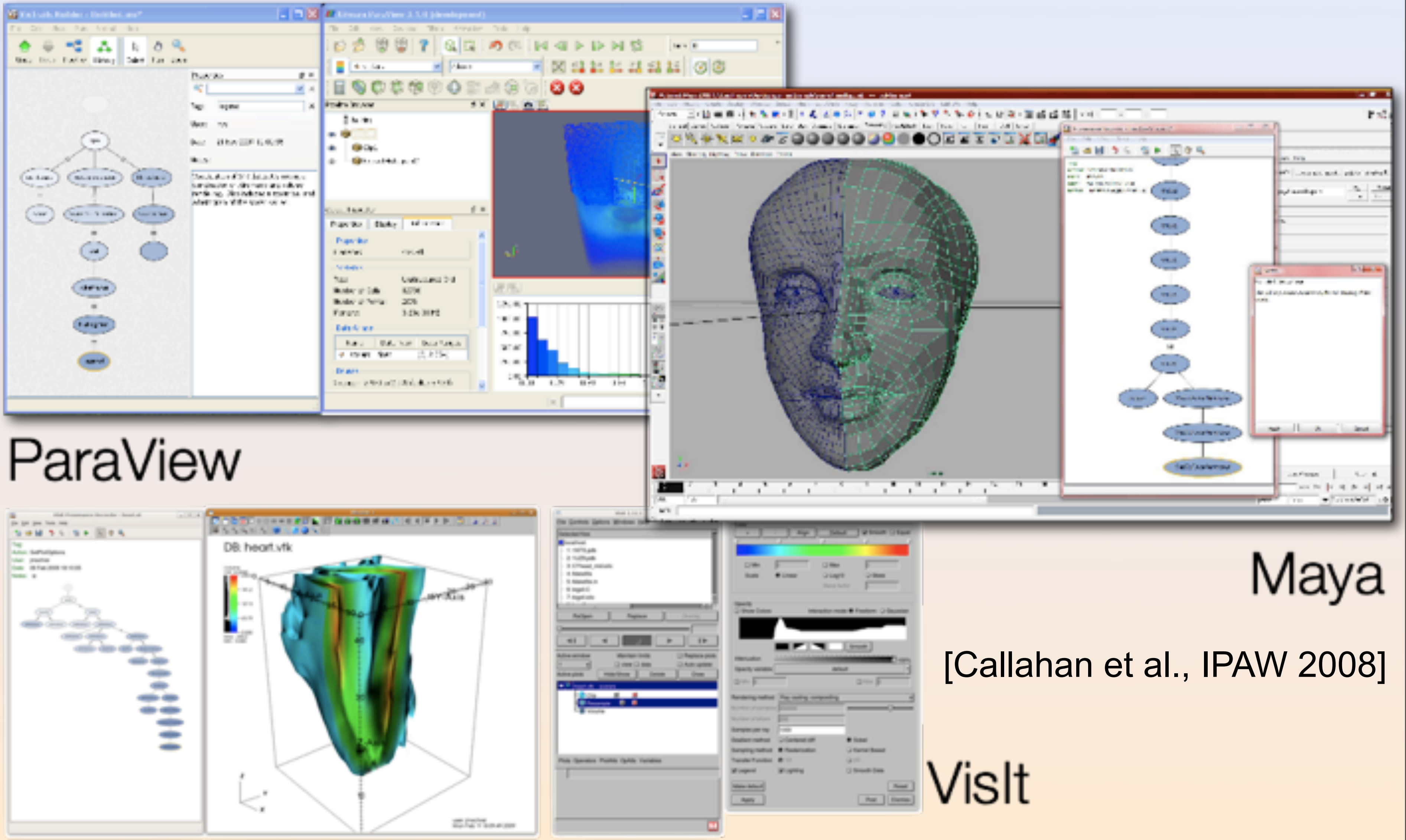


www.vistrails.org

Provenance "Plug-ins"



Provenance Enabling Tools



ParaView

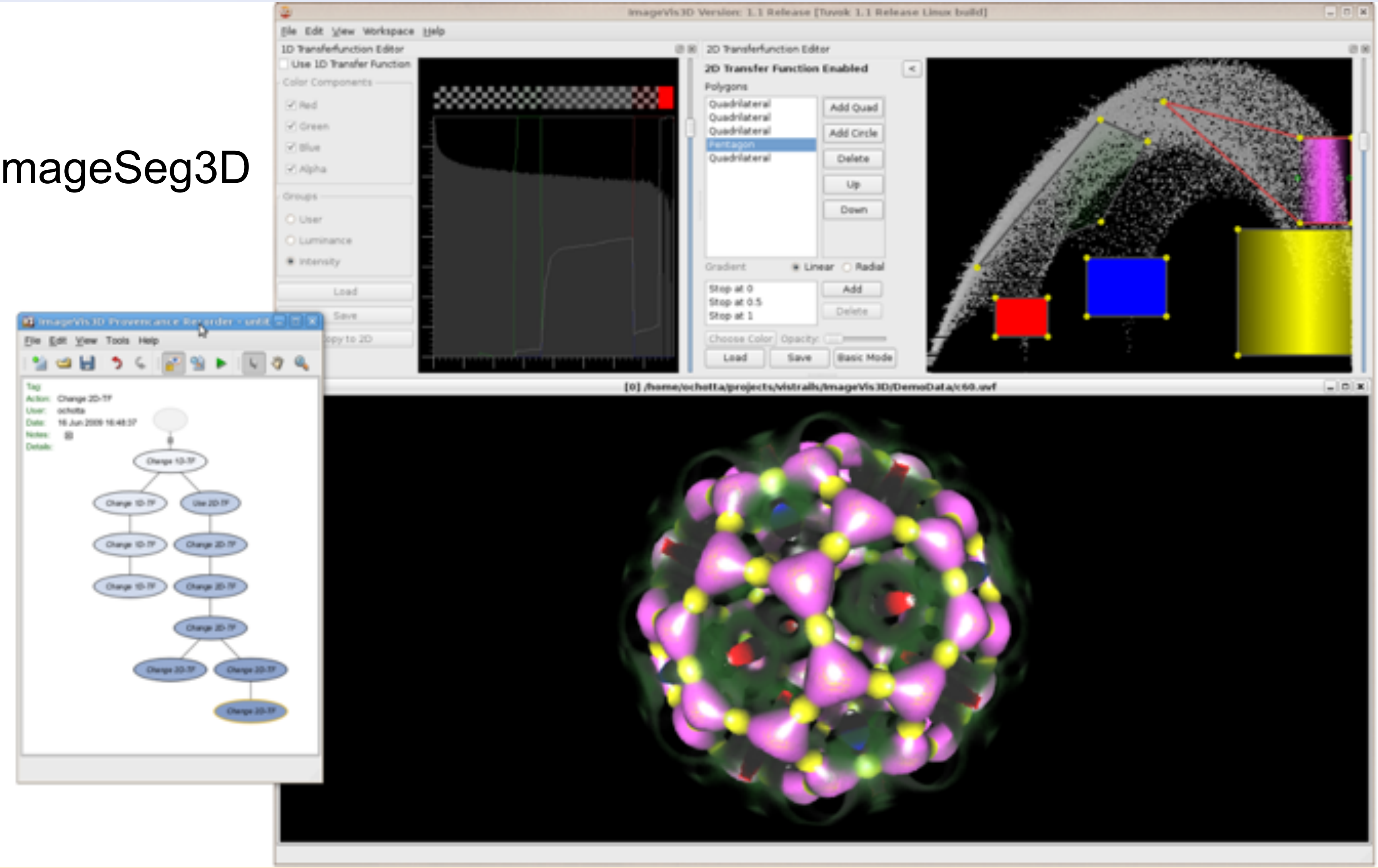
Maya

[Callahan et al., IPAW 2008]

VisIt

More plugins…

ImageSeg3D

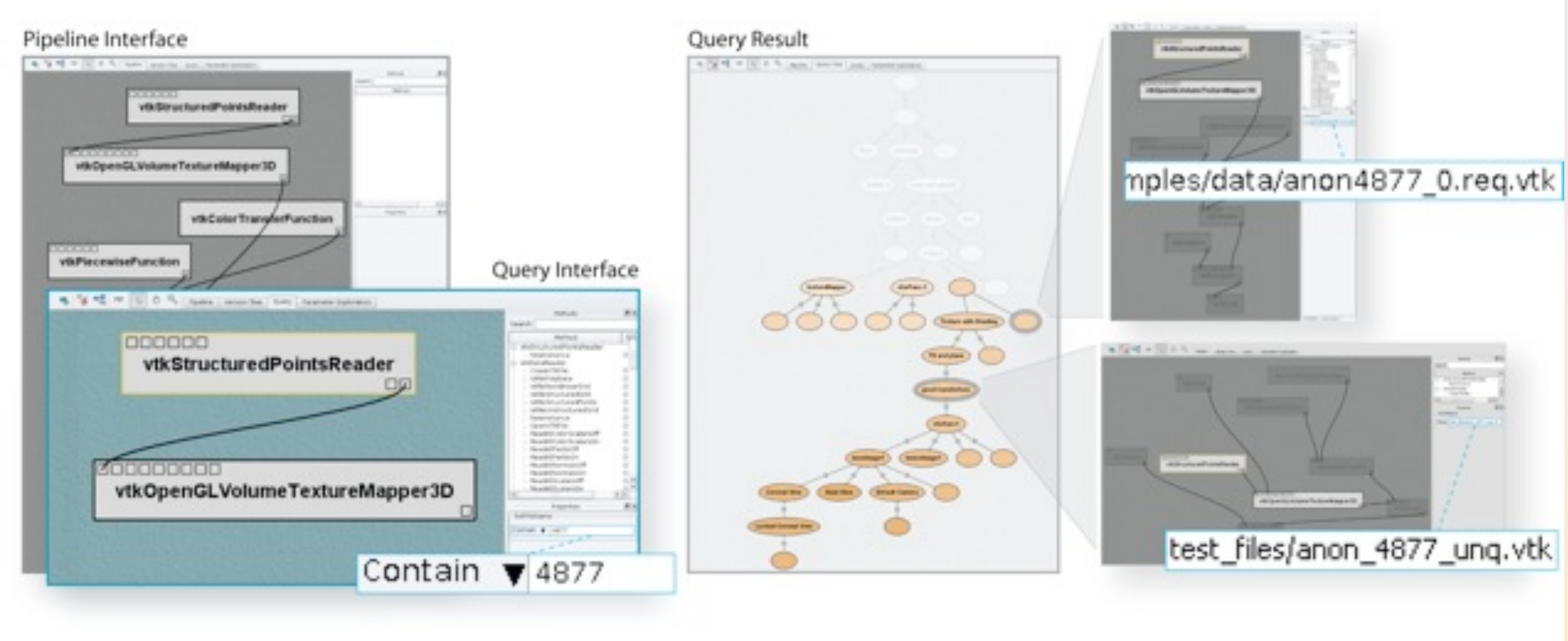


Tag:
Action: Change 2D-TF
User: ochotta
Date: 16 Jun 2009 16:48:37
Notes:
Details:

Sample of Ongoing Work

Querying Workflows by Example

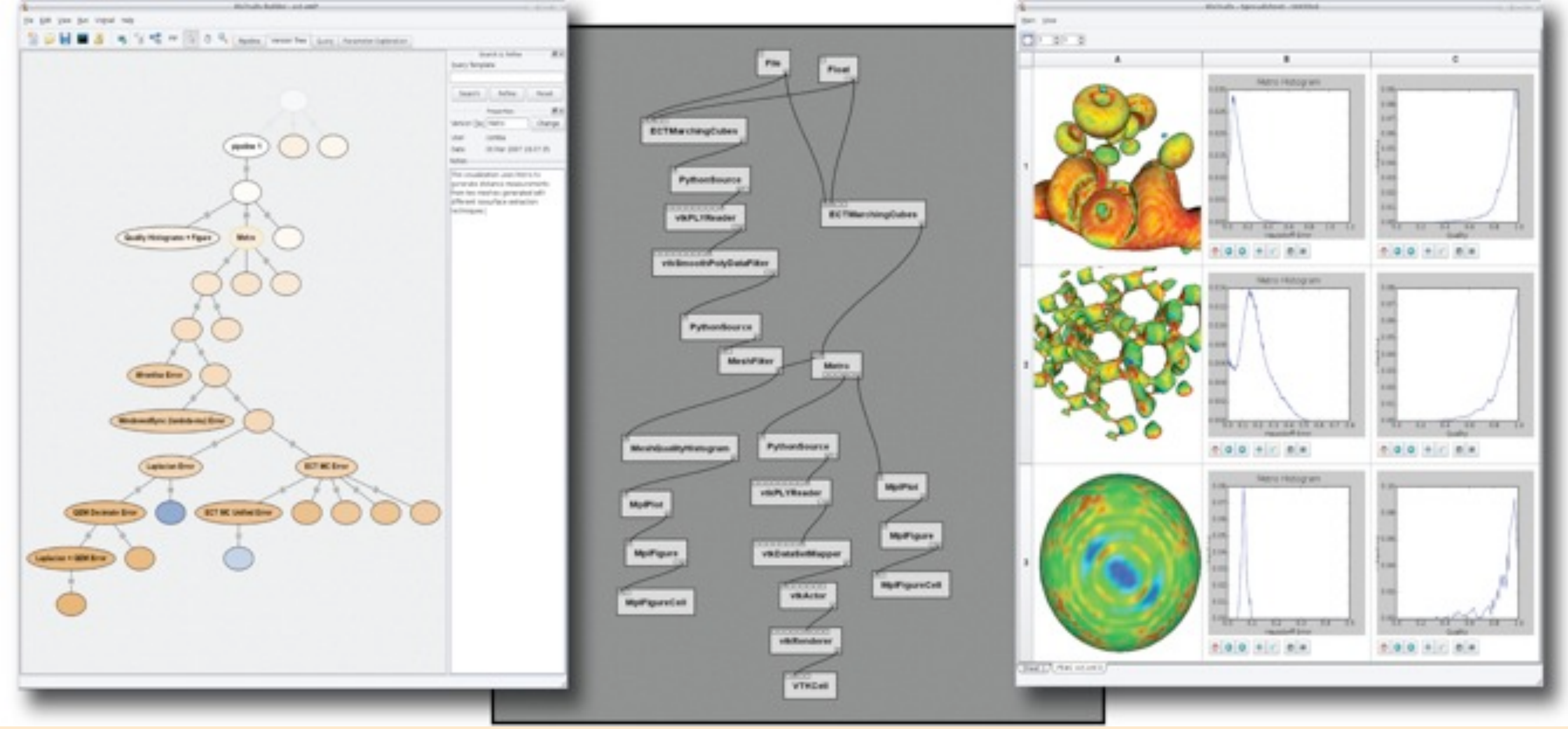
- Workflows are graphs: hard to specify queries using text!
- Querying workflows by example [Scheidegger et al., TVCG 2007; Beeri et al., VLDB 2006; Beeri et al. VLDB 2007]
  - WYSIWYQ -- What You See Is What You Query
  - Interface to create workflow is same as to query


Pipeline Interface
vtkStructuredPointsReader
vtkOpenGLVolumeTextureMapper3D
vtkColorTransferFunction
vtkPiecewiseFunction
Query Interface
vtkStructuredPointsReader
vtkOpenGLVolumeTextureMapper3D
Query Result
mples/data/anon4877_0.req.vtk
test_files/anon_4877_unq.vtk

Contain ▼ 4877

Creating Workflows

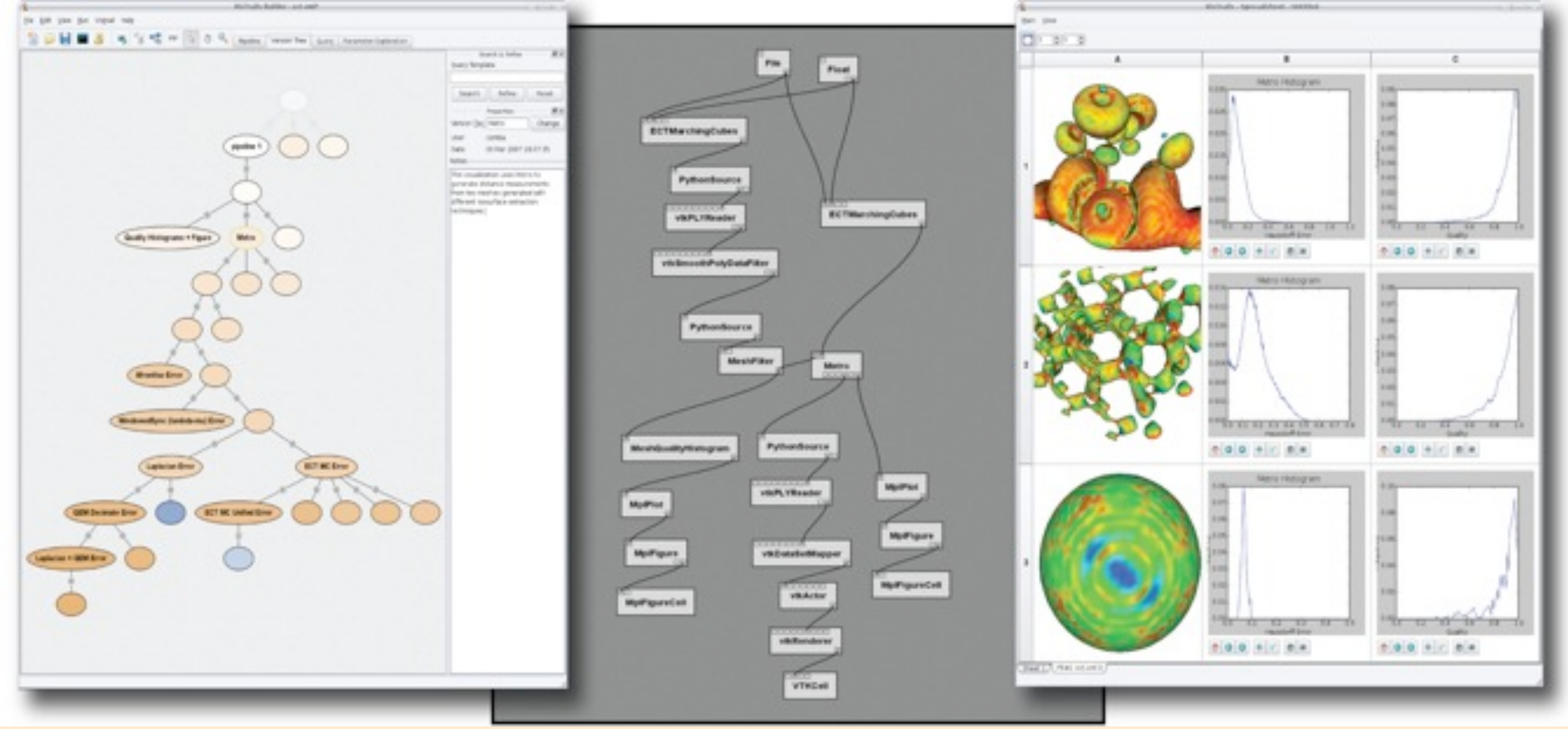
- Complex workflows are hard to create
  - Programming expertise
  - Domain knowledge
  - Familiarity with different tools



Creating Workflows

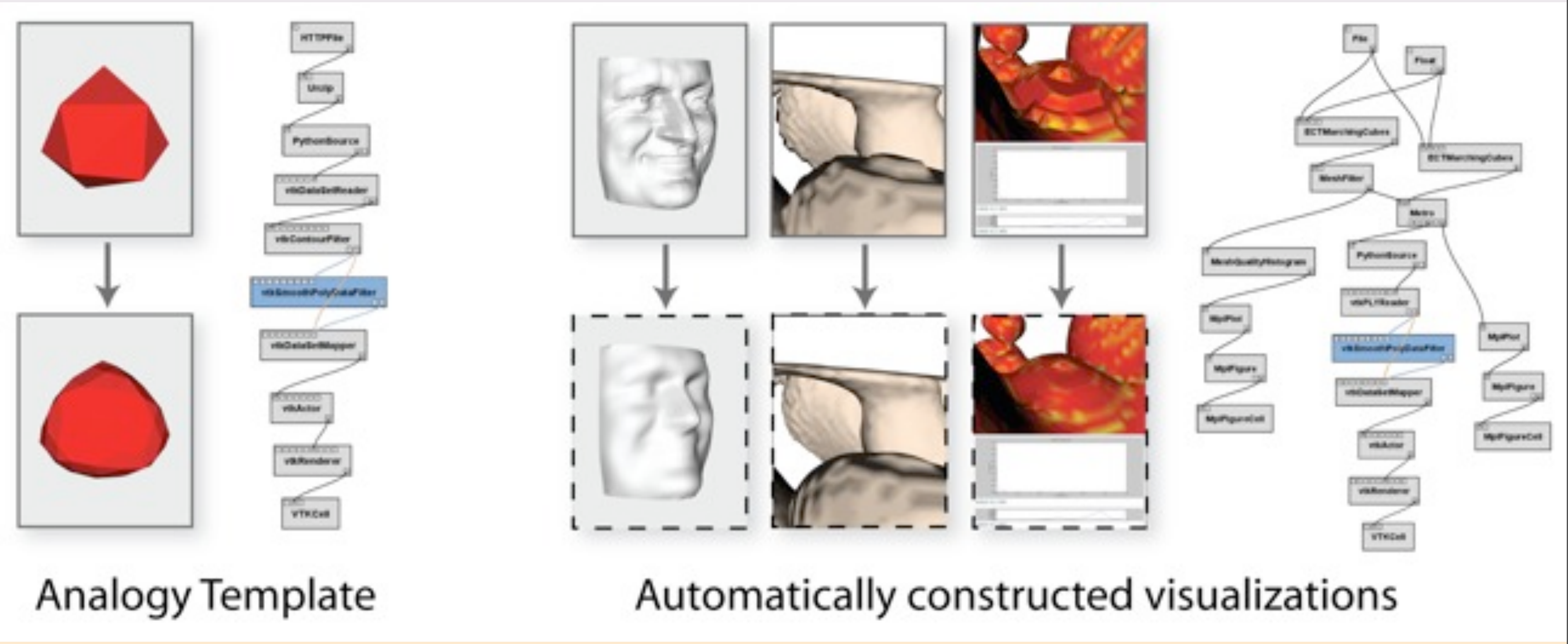
- Complex workflows are hard to create
  - Programming expertise
  - Domain knowledge
  - Familiarity with different tools

Steep learning curve



Creating Workflows by Analogy

- Use the wisdom of the crowds
  - Some workflow refinements are common, e.g., change the rendering technique, publish image on the Web
- Apply refinements by analogy, automatically [Scheidegger et al, IEEE TVCG 2007]

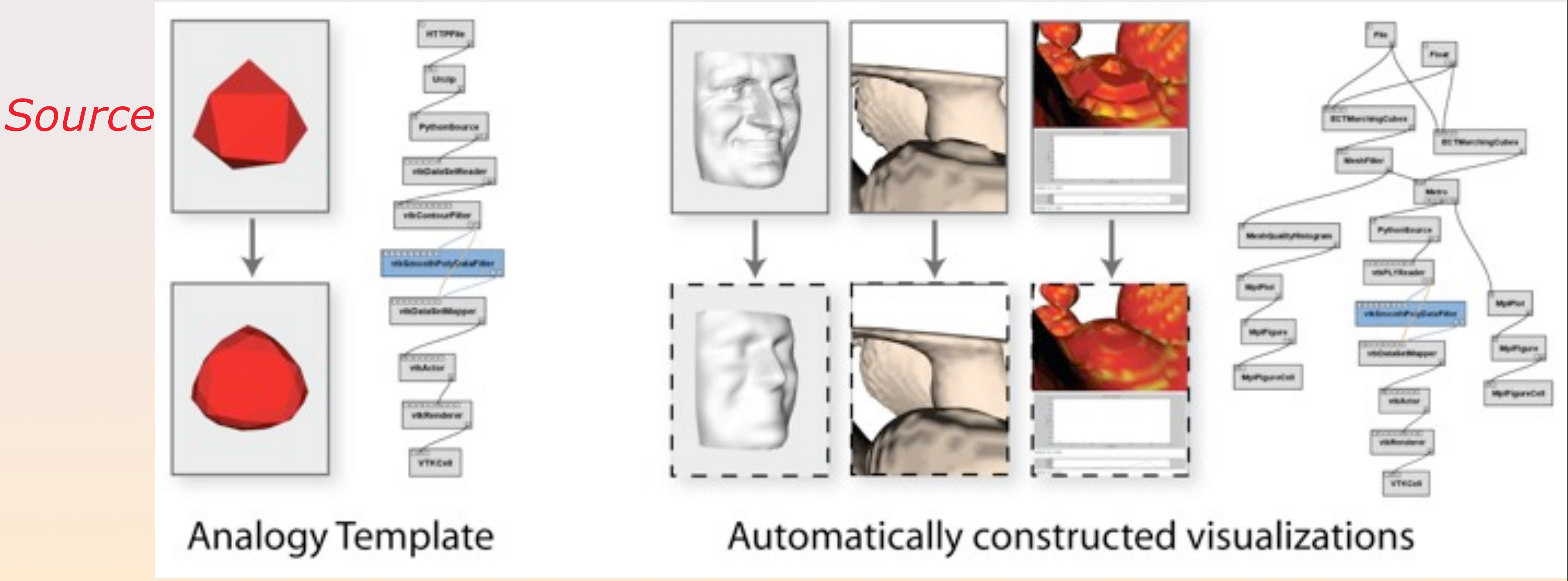


Analogy Template

Automatically constructed visualizations

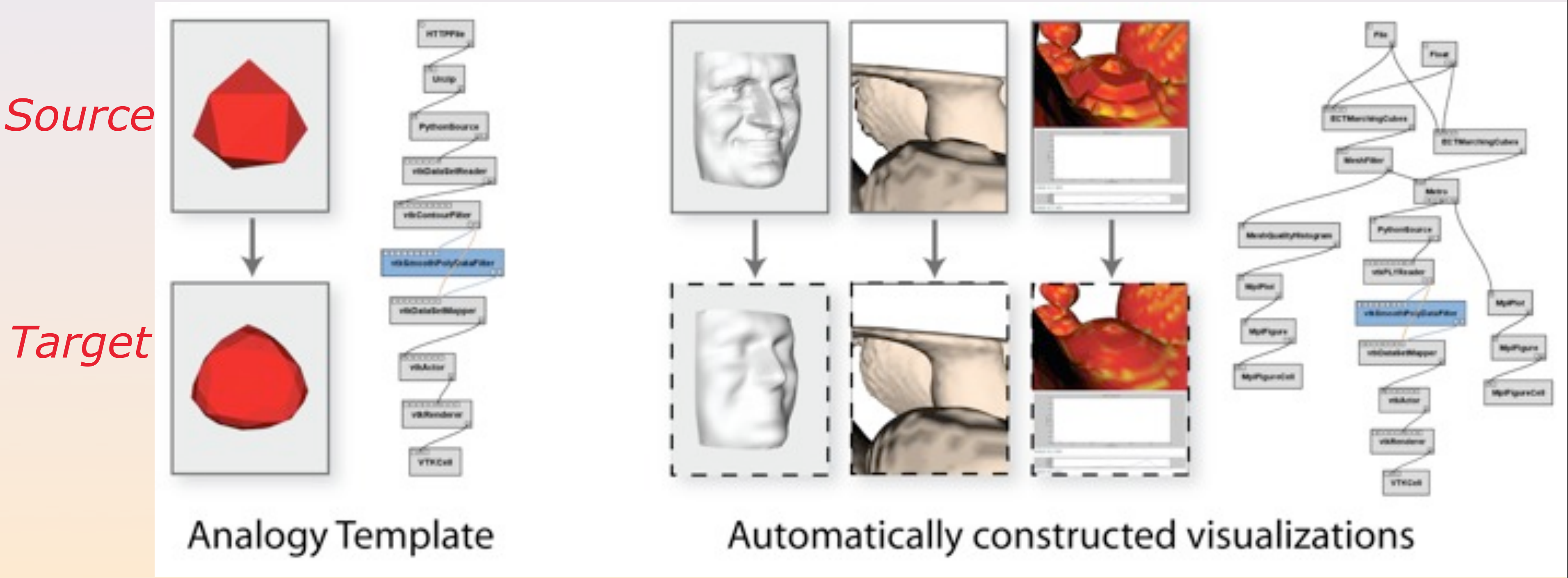
Creating Workflows by Analogy

- Use the wisdom of the crowds
  - Some workflow refinements are common, e.g., change the rendering technique, publish image on the Web
- Apply refinements by analogy, automatically [Scheidegger et al, IEEE TVCG 2007]

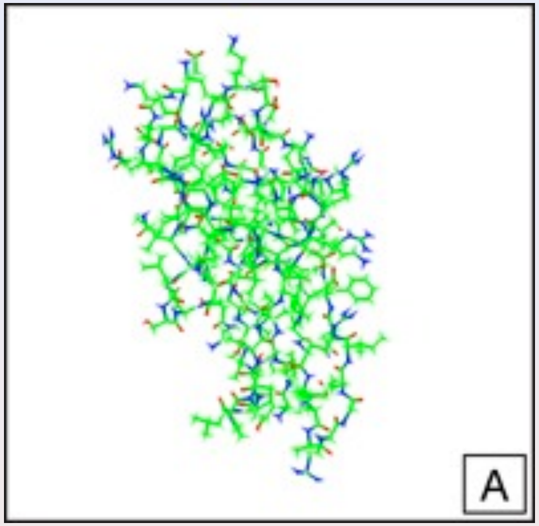


Creating Workflows by Analogy

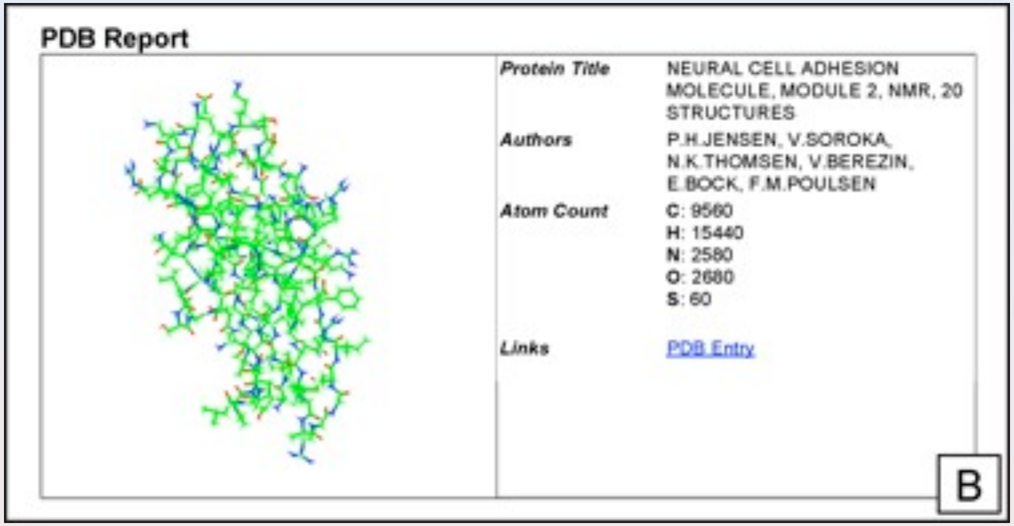
- Use the wisdom of the crowds
  - Some workflow refinements are common, e.g., change the rendering technique, publish image on the Web
- Apply refinements by analogy, automatically [Scheidegger et al, IEEE TVCG 2007]



Creating Workflows by Analogy



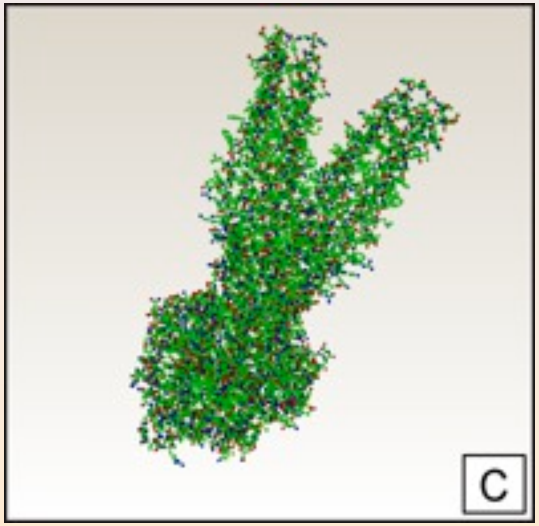
is to



PDB Report

Protein Title	NEURAL CELL ADHESION MOLECULE, MODULE 2, NMR, 20 STRUCTURES
Authors	P.H.JENSEN, V.SOROKA, N.K.THOMSEN, V.BEREZIN, E.BOCK, F.M.POULSEN
Atom Count	**C**: 9560 **H**: 15440 **N**: 2580 **O**: 2680 **S**: 60
Links	PDB Entry

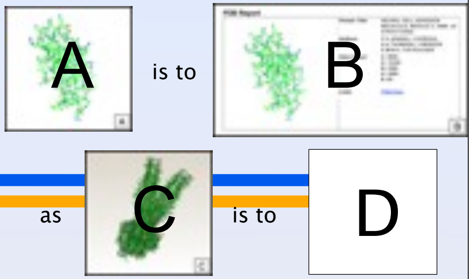
as



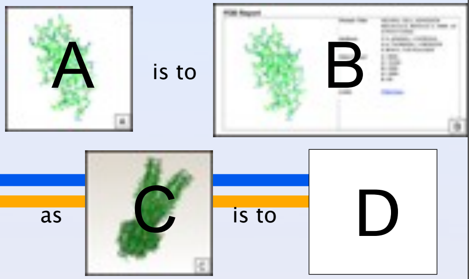
is to

?

Creating Workflows by Analogy



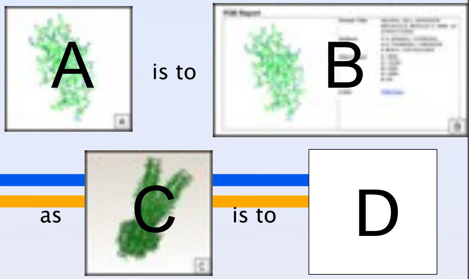
Creating Workflows by Analogy



1. Compute difference: Δ(A,B)
   - Just like a patch!
   - But…

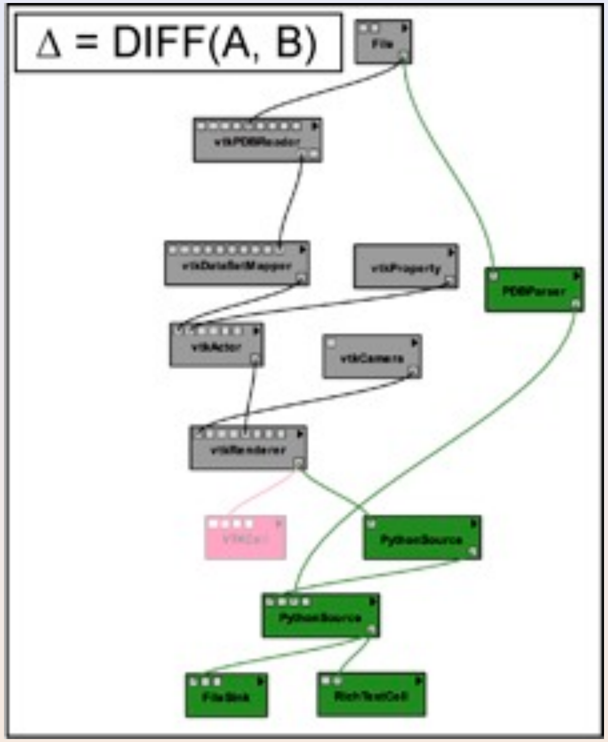
   D = Δ(A,B) ∘ C may not be a valid workflow

Creating Workflows by Analogy

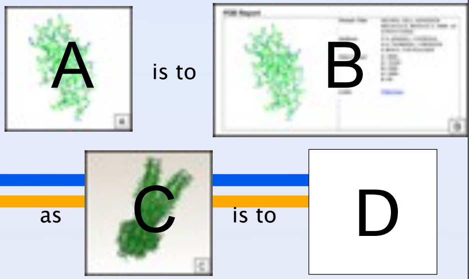


1. Compute difference: Δ(A,B)
   - Just like a patch!
   - But…

   D = Δ(A,B) ◦ C may not be a valid workflow



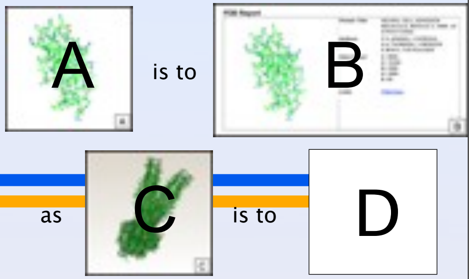
Creating Workflows by Analogy



1. Compute difference: Δ(A,B)
   - Just like a patch!
   - But…

   D = Δ(A,B) ◦ C may not be a valid workflow

Creating Workflows by Analogy

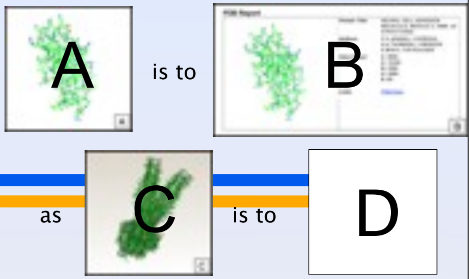


1. Compute difference: Δ(A,B)
   - Just like a patch!
   - But…

   D = Δ(A,B) ∘ C may not be a valid workflow

u Find correspondences between A and C: map(A,C)
   - Diffuse similarity scores across the product graph AxC using Eigenvalue decompositions

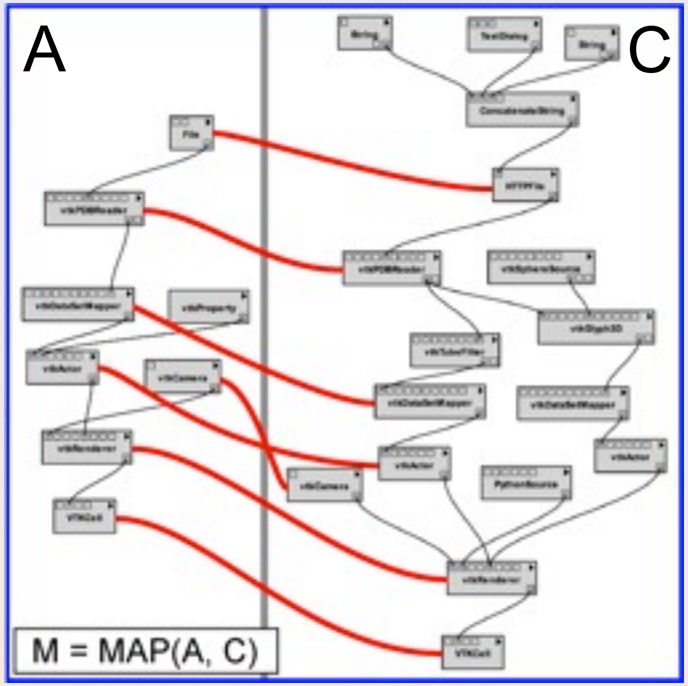
Creating Workflows by Analogy



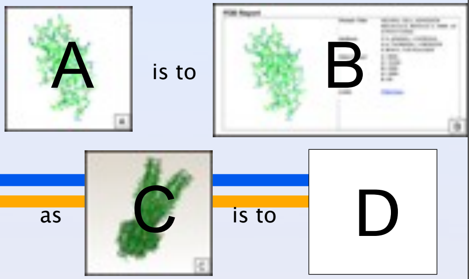
1. Compute difference: Δ(A,B)
   - Just like a patch!
   - But…

   D = Δ(A,B) ◦ C may not be a valid workflow

- Find correspondences between A and C: map(A,C)
  - Diffuse similarity scores across the product graph AxC using Eigenvalue decompositions



Creating Workflows by Analogy

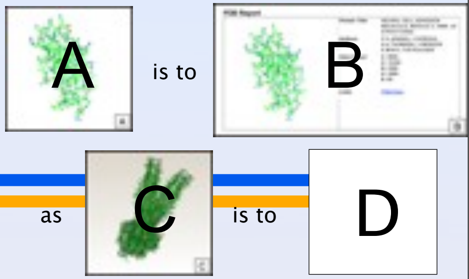


1. Compute difference: Δ(A,B)
   - Just like a patch!
   - But…

   D = Δ(A,B) ◦ C may not be a valid workflow

u Find correspondences between A and C: map(A,C)
   - Diffuse similarity scores across the product graph AxC using Eigenvalue decompositions

Creating Workflows by Analogy



1. Compute difference: Δ(A,B)
   - Just like a patch!
   - But…

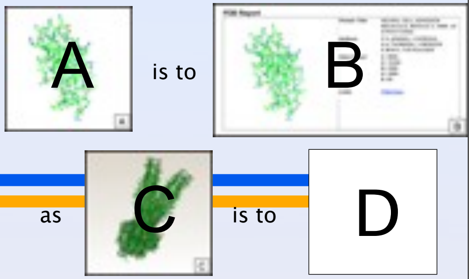
   D = Δ(A,B) ◦ C may not be a valid workflow

u Find correspondences between A and C: map(A,C)

- Diffuse similarity scores across the product graph AxC using Eigenvalue decompositions

u Compute mapped difference Δ_{AC}(A,B) =map(A,C) Δ(A,B)

Creating Workflows by Analogy



1. Compute difference: $\Delta(A,B)$
   - Just like a patch!
   - But…

   $D = \Delta(A,B) \circ C$ may not be a valid workflow

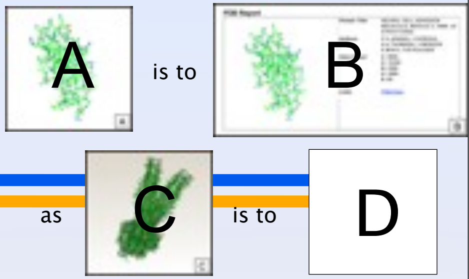
u Find correspondences between A and C: map(A,C)

- Diffuse similarity scores across the product graph AxC using Eigenvalue decompositions

u Compute mapped difference $\Delta_{AC}(A,B) = map(A,C)\ \Delta(A,B)$

u $D = \Delta_{AC}(A,B) \circ C$

Creating Workflows by Analogy



1. Compute difference: Δ(A,B)
   - Just like a patch!
   - But…

   D = Δ(A,B) ∘ C may not be a valid workflow

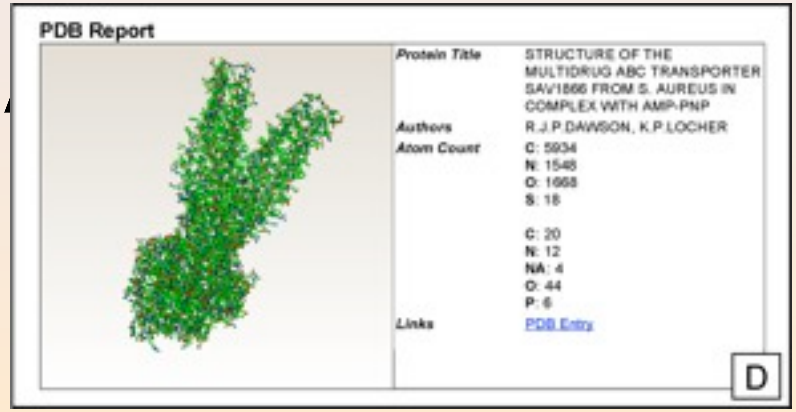
u Find correspondences between A and C: map(A,C)

   - Diffuse similarity scores across the product graph AxC using Eigenvalue decompositions

u Compute mapped difference Δ

(A,B) =map(A,C) Δ(A,B)

u D = Δ_{AC}(A,B) ∘ C

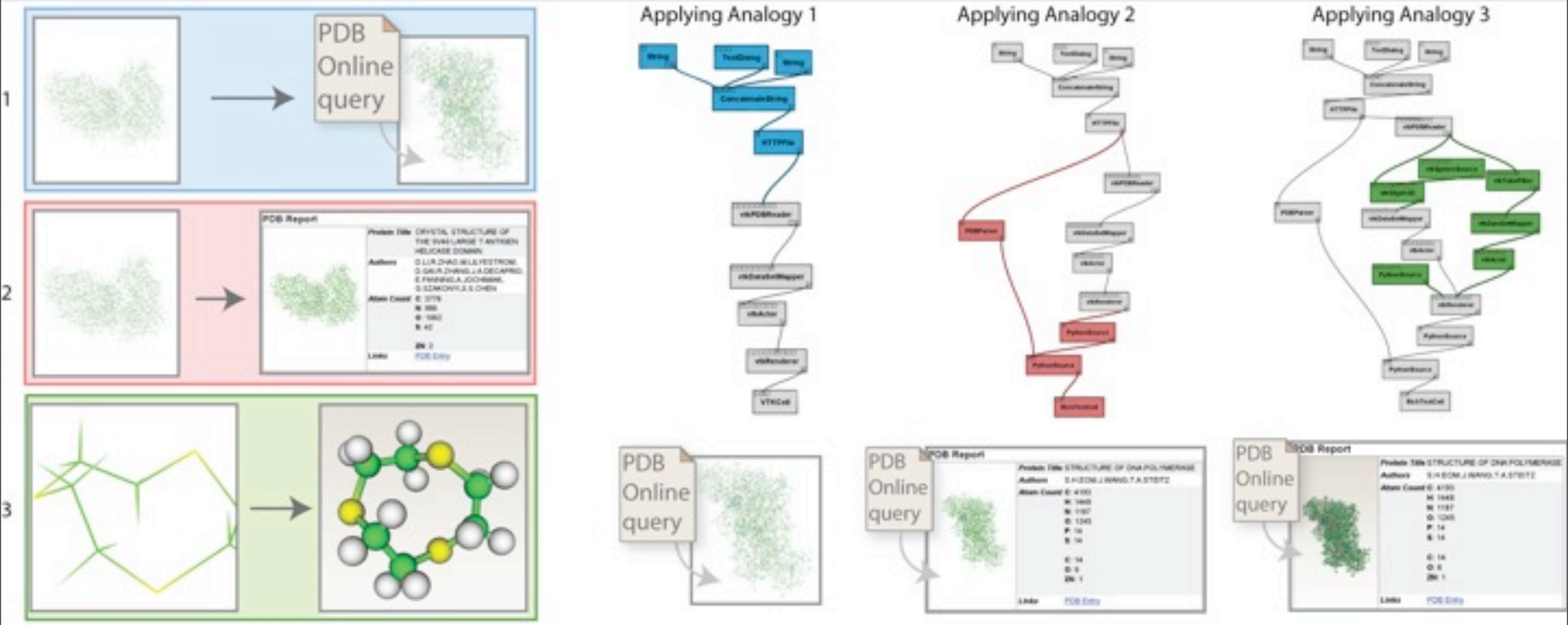


PDB Report

Protein Title	STRUCTURE OF THE MULTIDRUG ABC TRANSPORTER SAV1866 FROM S. AUREUS IN COMPLEX WITH AMP-PNP
Authors	R.J.P.DAWSON, K.P.LOCHER
Atom Count	C: 5934 N: 1548 O: 1668 S: 18 C: 20 N: 12 NA: 4 O: 44 P: 6
Links	PDB Entry

QBE and Analogies

- See paper:
  - Querying and Re-Using Workflows with VisTrails
    Carlos E. Scheidegger, David Koop, Huy Vo, Juliana Freire, and Claudio T. Silva **(Best Paper Award at VIS 2007)**

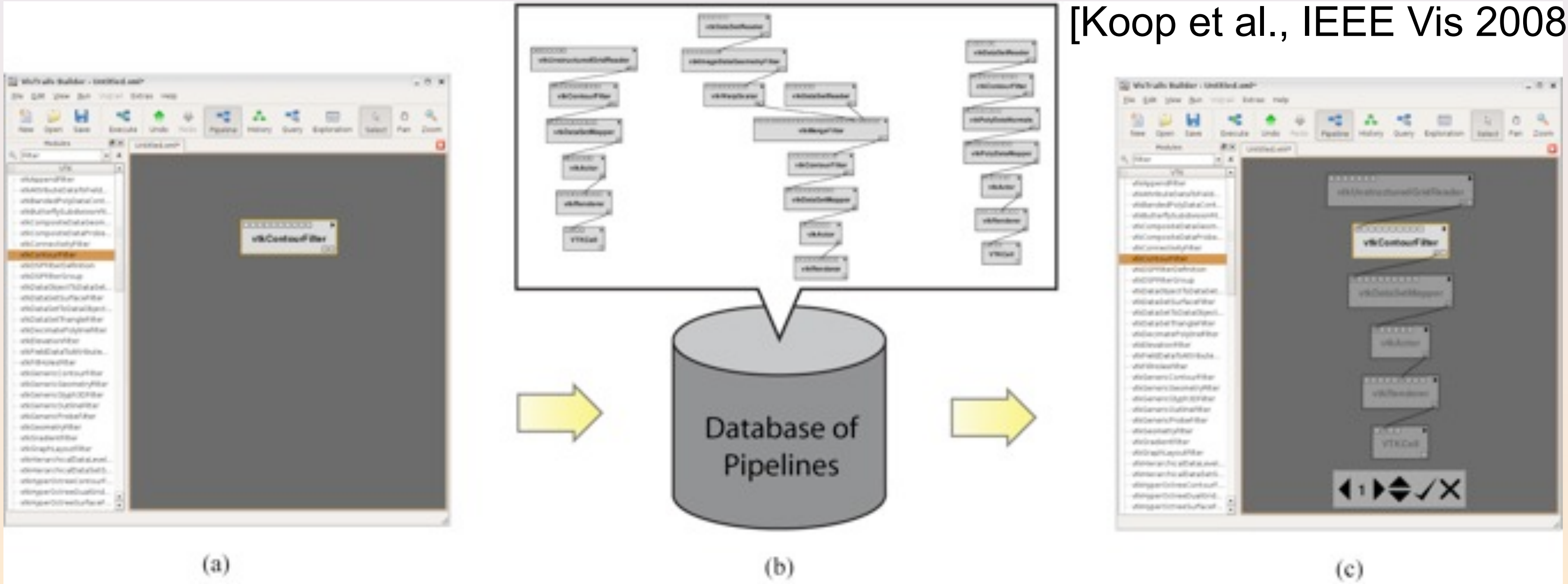


Analogy Templates

Automatically generated workflow sequence

VisComplete: A Workflow

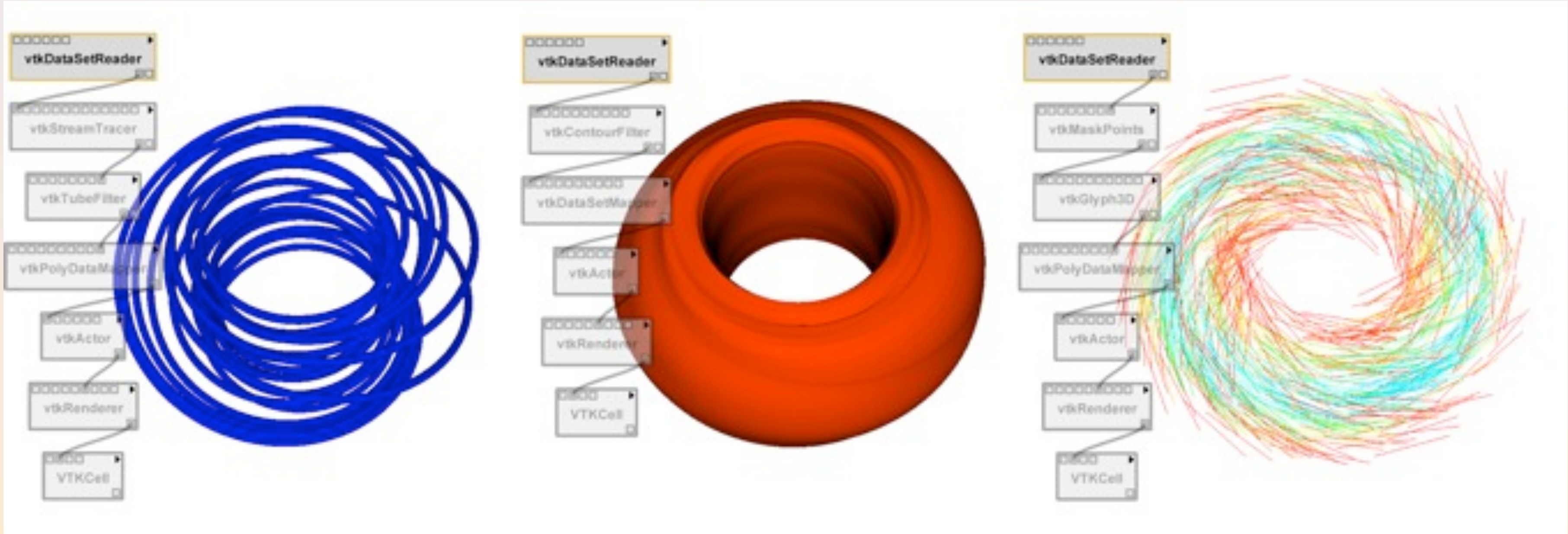
- Identify graph fragments that co-occur in a collection of workflows
- Predict sets of likely workflow additions to a given partial workflow



[Koop et al., IEEE Vis 2008]

VisComplete: A Workflow

- Similar to a Web browser suggesting URL completions
- Idea applicable to integration queries [Sarah Cohen-Boulakia et a., JBCB 2006; Talukdar et al., VLDB 2008]



VisComplete (video)

[Koop et al., IEEE Vis2008]

VisComplete (video)

[Koop et al., IEEE Vis2008]

VisComplete:
Data-driven Suggestions for
Visualization Systems

Acknowledgments: Funding

- This work is partially supported by the National Science Foundation, the Department of Energy, an IBM Faculty Award, and a University of Utah Seed Grant.

CMOP
Center for Coastal Margin Observation & Prediction

VACET
DOE SciDAC VISUALIZATION AND ANALYTICS CENTER FOR ENABLING TECHNOLOGIES

THE UNIVERSITY OF UTAH

National Science Foundation
WHERE DISCOVERIES BEGIN

SDM CENTER

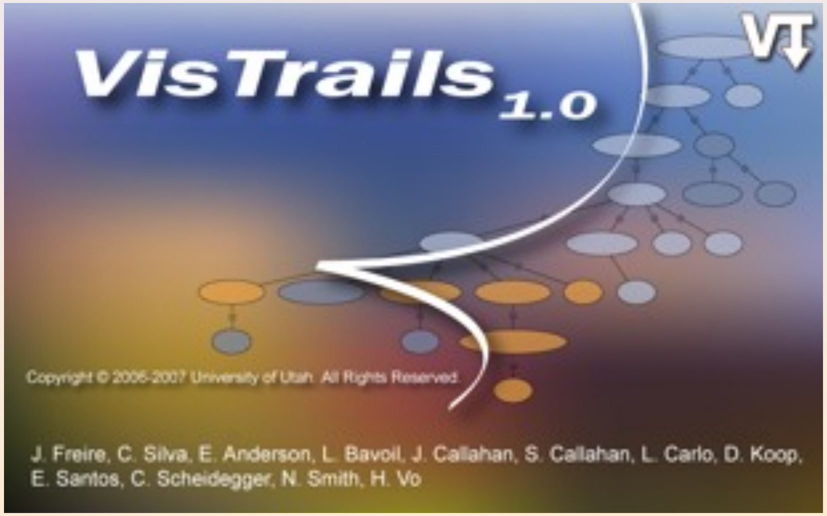
IBM

More info about VisTrails

google vistrails

Or

http://www.vistrails.org

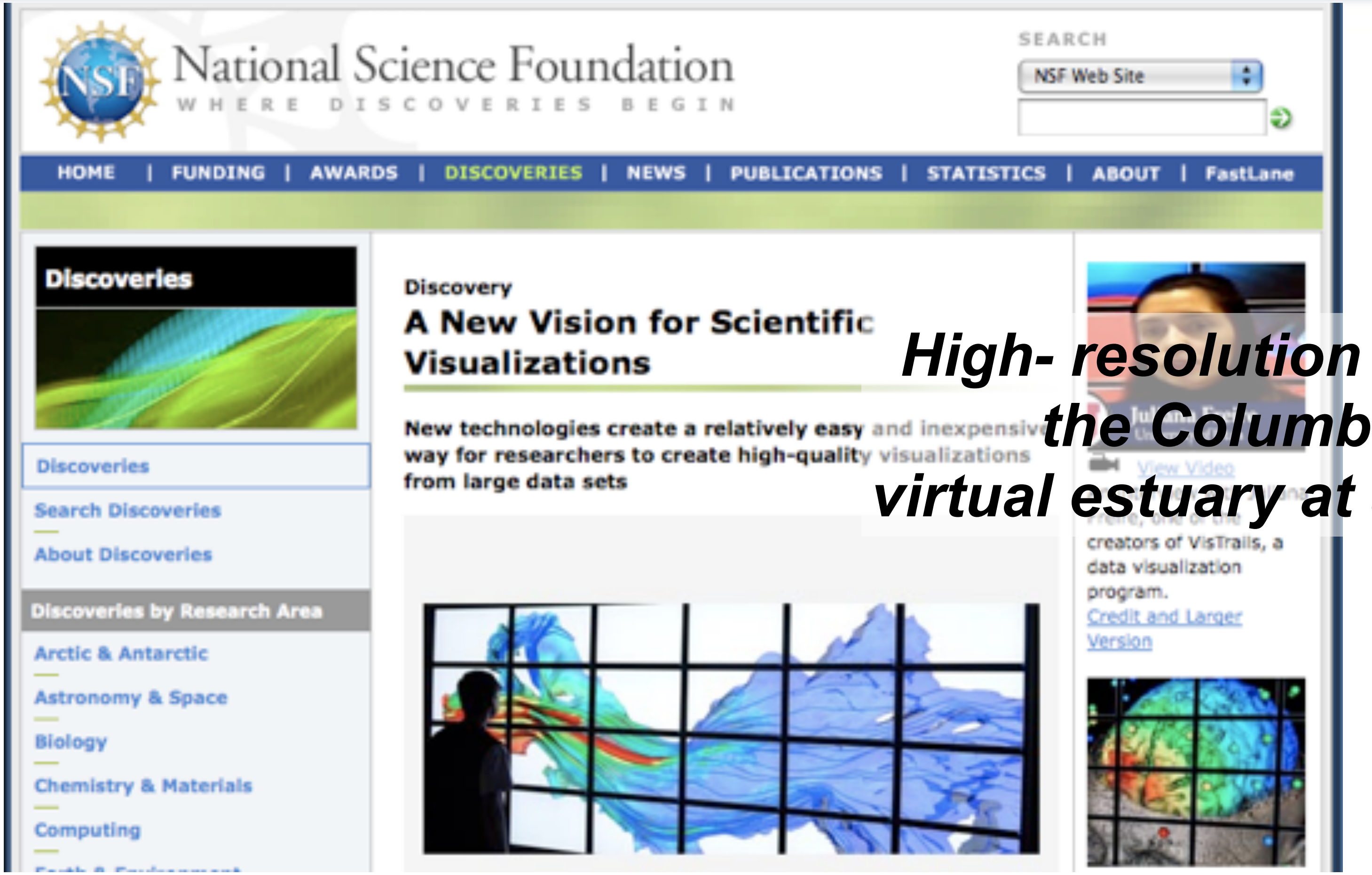


Copyright © 2006-2007 University of Utah. All Rights Reserved.

J. Freire, C. Silva, E. Anderson, L. Bavoil, J. Callahan, S. Callahan, L. Carlo, D. Koop,
E. Santos, C. Scheidegger, N. Smith, H. Vo

Emerging Work/ Applications

Visualization at large and on the go



Discovery

A New Vision for Scientific Visualizations

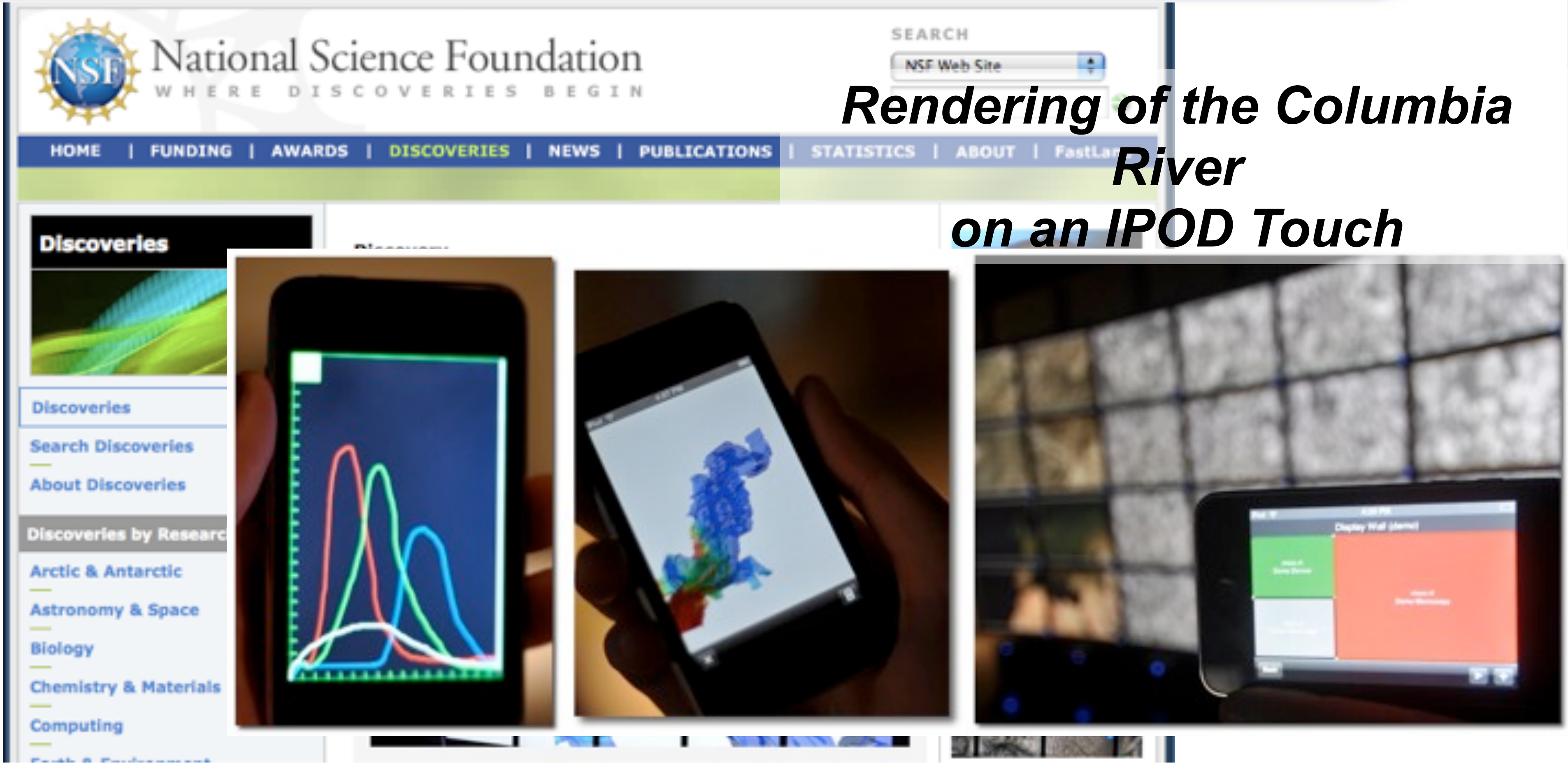
New technologies create a relatively easy and inexpensive way for researchers to create high-quality visualizations from large data sets

creators of VisTrails, a data visualization program.

Credit and Larger Version

High- resolution rendering of the Columbia River virtual estuary at a display wall

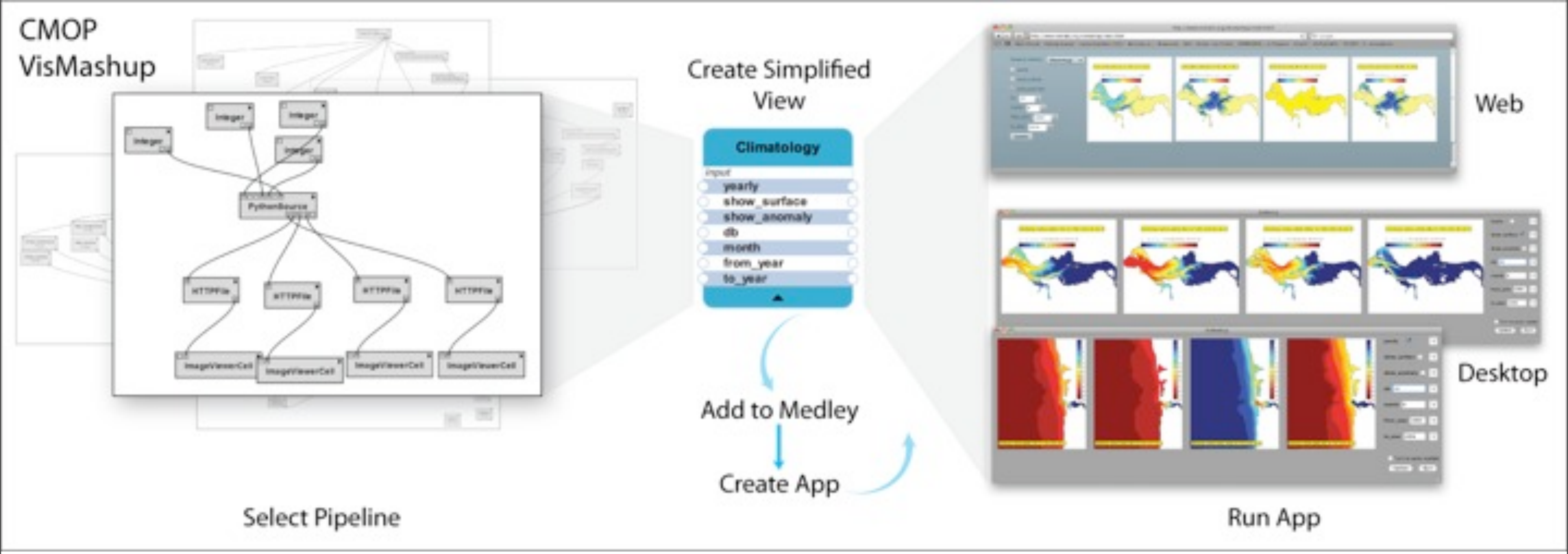
Visualization at large and on the go



Rendering of the Columbia River on an IPOD Touch

VisTrails: Science Dissemination

- Science mashups: simplify data exploration through visualization



[Santos et al, IEEE Vis 2009 (cond. accept)]

Scientific Publications and Provenance

J Appl Physiol 98: 2191–2196, 2005.
First published March 17, 2005; doi:10.1152/japplphysiol.00216.2005.

Improved muscular efficiency displayed as Tour de France champion matures

Edward F. Coyle
Human Performance Laboratory, Department of Kinesiology and Health Education, The University of Texas at Austin, Austin, Texas

Submitted 22 February 2005; accepted in final form 10 March 2005

Coyle, Edward F. Improved muscular efficiency displayed as Tour de France champion matures. *J Appl Physiol* 98: 2191–2196, 2005. First published March 17, 2005;doi:10.1152/japplphysiol.00216.2005.—This case describes the physiological maturation from ages 21 to 28 yr of the bicyclist who has now become the six-time consecutive Grand Champion of the Tour de France, at ages 27–32 yr. Maximal oxygen uptake ($\dot{V}O_{2\,max}$) in the trained state remained at ~6 l/min, lean body weight remained at ~70 kg, and maximal heart rate declined from 207 to 200 beats/min. Blood lactate threshold was typical of competitive cyclists in that it occurred at 76–85% $\dot{V}O_{2\,max}$, yet maximal blood lactate concentration was remarkably low in the trained state. It appears that an 8% improvement in muscular efficiency and thus power production when cycling at a given oxygen uptake ($\dot{V}O_2$) is the characteristic that improved most as this athlete matured from ages 21 to 28 yr. It is noteworthy that at age 25 yr, this champion developed advanced cancer, requiring surgeries and chemotherapy. During the months leading up to each of his Tour de France victories, he reduced body weight and body fat by 4–7 kg (i.e., ~7%). Therefore, over the 7-yr period, an improvement in muscular efficiency and reduced body fat contributed equally to a remarkable 18% improvement in his steady-state power per kilogram body weight when cycling at a given $\dot{V}O_2$ (e.g., 5 l/min). It is hypothesized that the improved muscular efficiency probably reflects changes in muscle myosin type stimulated from years of training intensely for 3–6 h on most days.

maximum oxygen uptake; blood lactate concentration

MUCH HAS BEEN LEARNED about the physiological factors that contribute to endurance performance ability by simply describing the characteristics of elite endurance athletes in sports such as distance running, bicycle racing, and cross-country skiing. The numerous physiological determinants of endurance have been organized into a model that integrates such factors as maximal oxygen uptake ($\dot{V}O_{2\,max}$), the blood lactate threshold, and muscular efficiency, as these have been found to be the most important variables (7, 8, 15, 21). A common approach has been to measure these physiological factors in a given athlete at one point in time during their competitive career and to compare this individual's profile with that of a population of peers (4, 6, 15, 16, 21). Although this approach describes the variations that exist within a population, it does not provide information about the extent to which a given athlete can improve their specific physiological determinants of endurance with years of continued training as the athlete matures and reaches his/her physiological potential. There are remarkably few longitudinal reports documenting the changes in physiological factors that accompany years of continued endurance training at the level performed by elite endurance athletes.

This case study reports the physiological changes that occur in an individual bicycle racer during a 7-yr period spanning ages 21 to 28 y. Description of this person is noteworthy for two reasons. First, he rose to become a six-time and present Grand Champion of the Tour de France, and thus adaptations relevant to this feat were identified. Remarkably, he accomplished this after developing and receiving treatment for advanced cancer. Therefore, this report is also important because it provides insight, although limited, regarding the recovery of "performance physiology" after successful treatment for advanced cancer. The approach of this study will be to report results from standardized laboratory testing on this individual at five time points corresponding to ages 21.1, 21.5, 22.0, 25.8, and 28.2 yr.

METHODS

General testing sequence. On reporting to the laboratory, training, racing, and medical histories were obtained, body weight was measured (±0.1 kg), and the following tests were performed after informed consent was obtained, with procedures approved by the Internal Review Board of The University of Texas at Austin. Mechanical efficiency and the blood lactate threshold (LT) were determined as the subject bicycled a stationary ergometer for 25 min, with work rate increasing progressively every 5 min over a range of 50, 60, 70, 80, and 90% $\dot{V}O_{2\,max}$. After a 10- to 20-min period of active recovery, $\dot{V}O_{2\,max}$ when cycling was measured. Thereafter, body composition was determined by hydrostatic weighing and/or analysis of skin-fold thickness (34, 35).

Measurement of $\dot{V}O_{2\,max}$. The same Monark ergometer (model 819) equipped with a racing seat and drop handlebars and pedals for cycling shoes was used for all cycle testing, and seat height and saddle position were held constant. The pedal's crank length was 170 mm. $\dot{V}O_{2\,max}$ was measured during continuous cycling lasting between 8 and 12 min, with work rate increasing every 2 min. A leveling off of oxygen uptake ($\dot{V}O_2$) always occurred, and this individual cycled until exhaustion at a final power output that was 10–20% higher than the minimal power output needed to elicit $\dot{V}O_{2\,max}$. A venous blood sample was obtained 3–4 min after exhaustion for determination of blood lactate concentration after maximal exercise, as described below. The subject breathed through a Daniels valve; expired gases were continuously sampled from a mixing chamber and analyzed for O_2 (Applied Electrochemistry S3A) and CO_2 (Beckman LB-2). Inspired air volumes were measured using a dry-gas meter (Parkinson-Cowan CD4). These instruments were interfaced with a computer that calculated $\dot{V}O_2$ every 30 s. The same equipment for indirect calorimetry was used over the 7-yr period, with gas analyzers calibrated against the same known gases and the dry-gas meter calibrated periodically to a 350-liter Tissot spirometer.

Blood LT. The subject pedaled the Monark ergometer (model 819) continuously for 25 min at work rates eliciting ~50, 60, 70, 80, and 90% $\dot{V}O_{2\,max}$ for each successive 5-min stage. The calibrated ergometer was set in the constant power mode, and the subject maintained a pedaling cadence of 85 rpm. Blood samples were obtained either from

Address for reprint requests and other correspondence: E. F. Coyle, Bellmont Hall 222, Dept. of Kinesiology and Health Education, The Univ. of Texas at Austin, Austin, TX 78712 (E-mail: coyle@mail.utexas.edu).

The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

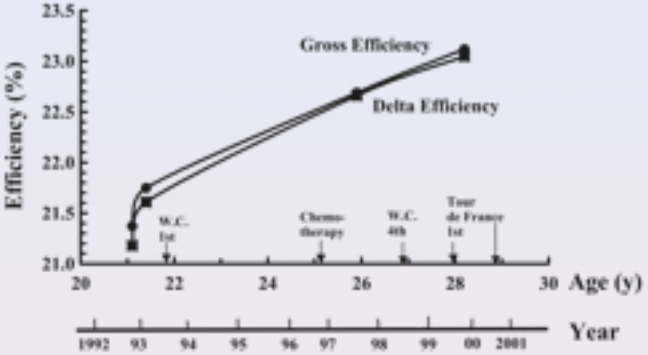


Fig. 1. Mechanical efficiency when bicycling expressed as "gross efficiency" and "delta efficiency" over the 7-yr period in this individual. WC, World Bicycle Road Racing Championships, 1st and 4th place, respectively. Tour de France 1st, Grand Champion of the Tour de France in 1999–2004.

METHODS

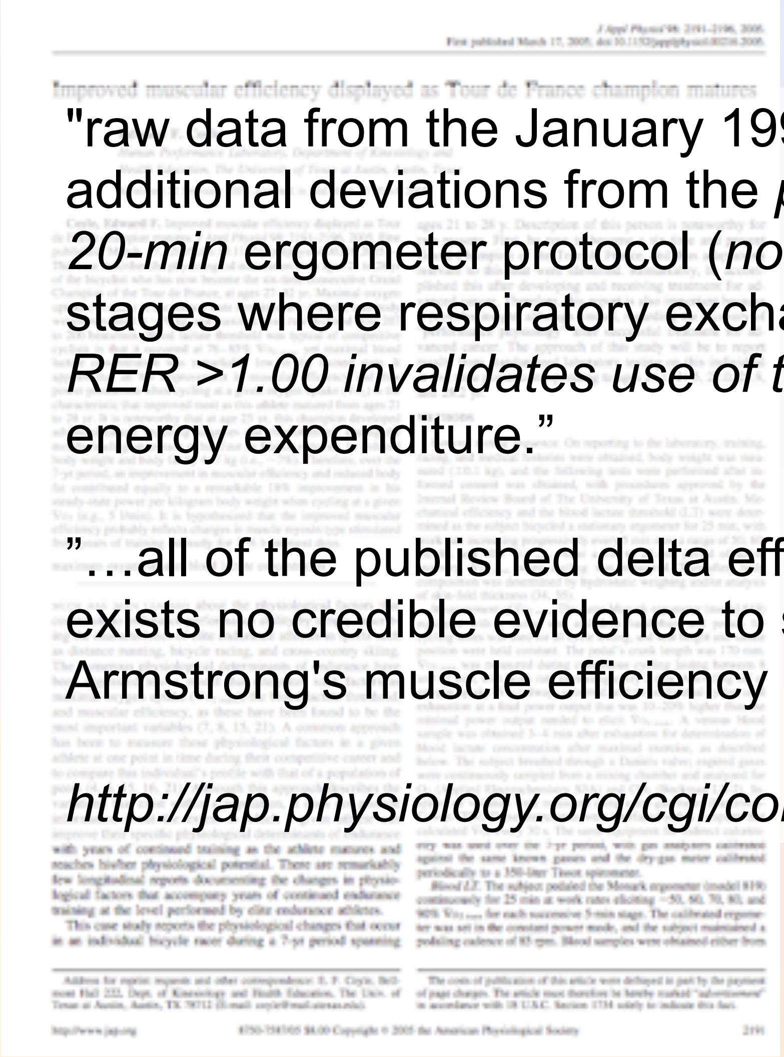
General testing sequence. On reporting to the laboratory, training, racing, and medical histories were obtained, body weight was measured (±0.1 kg), and the following tests were performed after informed consent was obtained, with procedures approved by the Internal Review Board of The University of Texas at Austin. Mechanical efficiency and the blood lactate threshold (LT) were determined as the subject bicycled a stationary ergometer for 25 min, with work rate increasing progressively every 5 min over a range of 50, 60, 70, 80, and 90% $\dot{V}O_{2\,max}$. After a 10- to 20-min period of active recovery, $\dot{V}O_{2\,max}$ when cycling was measured. Thereafter, body composition was determined by hydrostatic weighing and/or analysis of skin-fold thickness (34, 35).

Scientific Publications and Provenance

"raw data from the January 1993 test that revealed several additional deviations from the *published* methodology. Coyle used a *20-min* ergometer protocol (*not 25 min*), including 2- and 3-min stages where respiratory exchange ratios (RER) exceeded 1.00. An *RER >1.00 invalidates use of the Lusk equations* (5) to estimate energy expenditure."

"…all of the published delta efficiency values are wrong. …there exists no credible evidence to support Coyle's conclusion that Armstrong's muscle efficiency improved."

http://jap.physiology.org/cgi/content/full/105/3/1020



J Appl Physiol 98: 2191–2196, 2005.
First published March 17, 2005; doi:10.1152/japplphysiol.00216.2005.

Improved muscular efficiency displayed as Tour de France champion matures

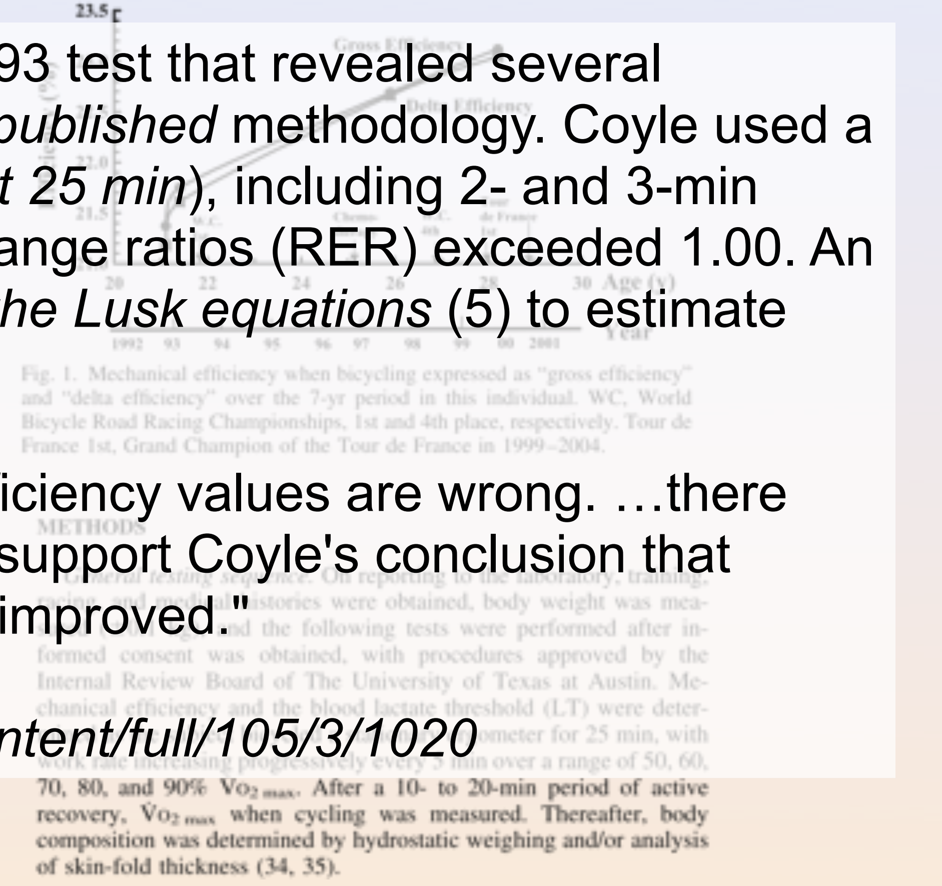


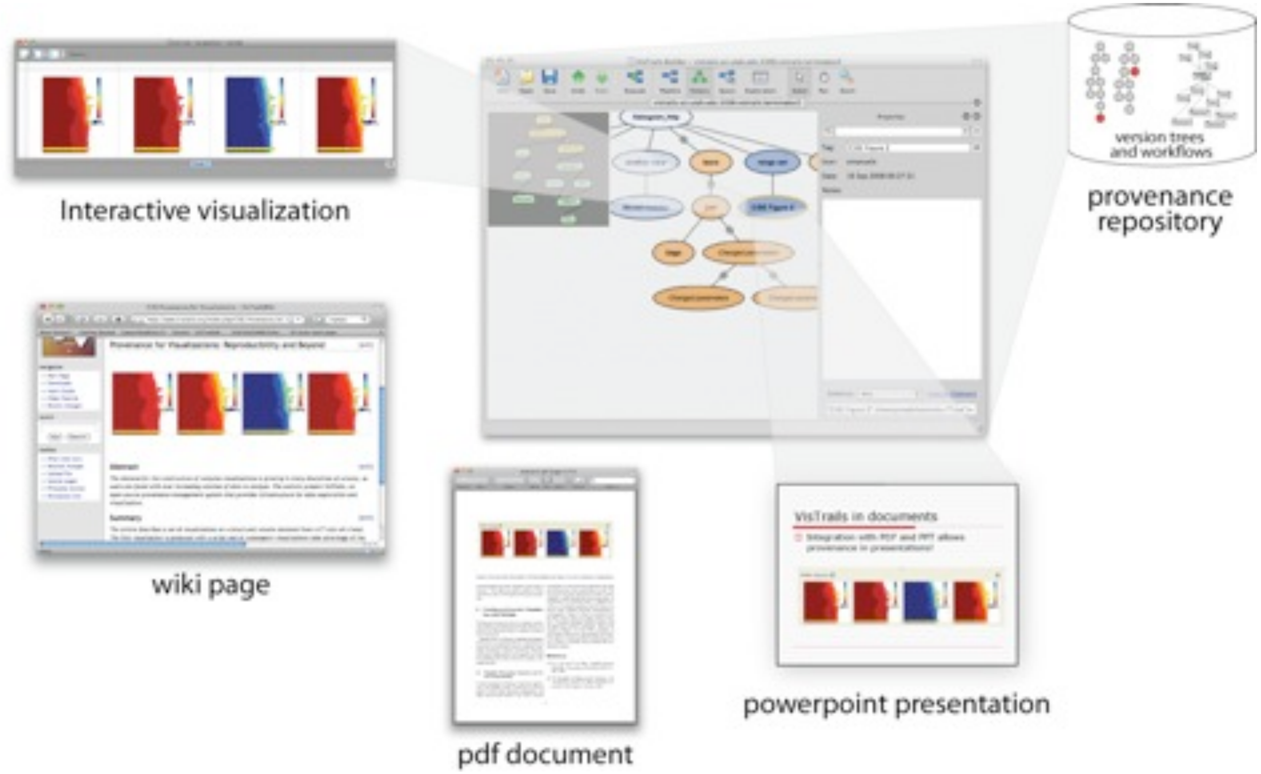
Fig. 1. Mechanical efficiency when bicycling expressed as "gross efficiency" and "delta efficiency" over the 7-yr period in this individual. WC, World Bicycle Road Racing Championships, 1st and 4th place, respectively. Tour de France 1st, Grand Champion of the Tour de France in 1999–2004.

METHODS

70, 80, and 90% $\dot{V}O_{2\,max}$. After a 10- to 20-min period of active recovery, $\dot{V}O_{2\,max}$ when cycling was measured. Thereafter, body composition was determined by hydrostatic weighing and/or analysis of skin-fold thickness (34, 35).

VisTrails: Science Dissemination

- Provenance-rich documents and publications


Interactive visualization
version trees and workflows
provenance repository
wiki page
Provenance for Visualizations: Reproducibility and Beyond
Abstract
Summary
pdf document
VisTrails in documents
Integration with PDF and PPT allows provenance in presentations!
powerpoint presentation

The Provenance-Rich Paper

Provenance and Teaching (1)

- Leverage provenance to improve the way we teach CS and Science
  - http://www.vistrails.org/index.php/SciVisFall2008
  - Lecture provenance: student can reproduce results

Provenance and Teaching (1)

- Leverage provenance to improve the way we teach CS and Science
  - http://www.vistrails.org/index.php/SciVisFall2008
  - Lecture provenance: student can reproduce results

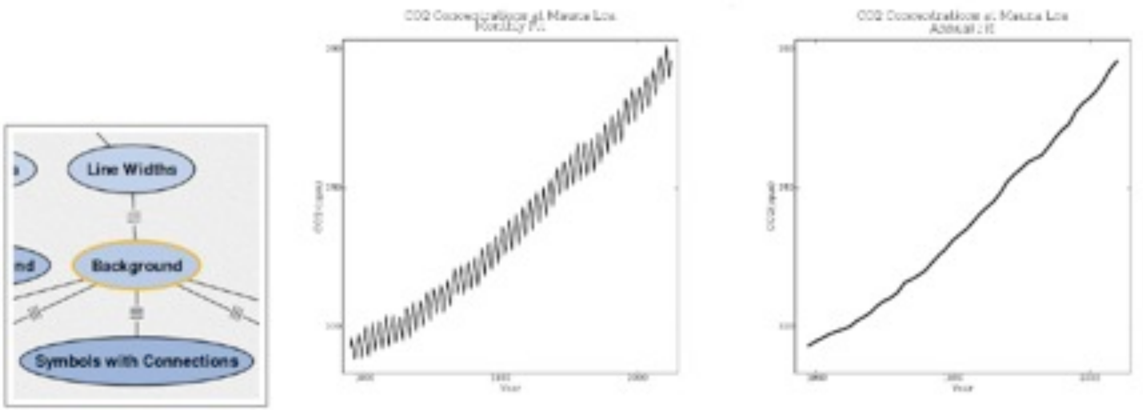


Figure 5.2: Plots of the Mauna Loa data set showing monthly measurements (left) with the yearly trend (right) using the principles for improving vision. The plot on the right is the same that was shown previously in Figure 5.1.

Provenance and Teaching (2)

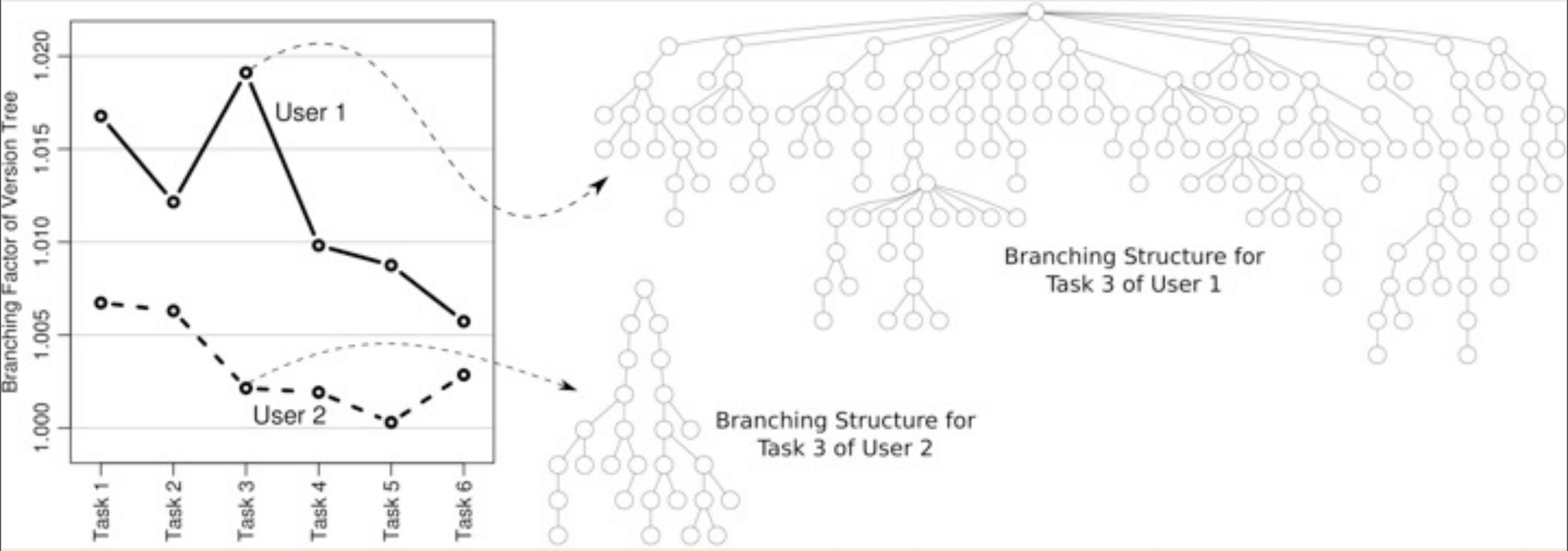
Provenance and Teaching (2)

- Homework provenance provides insights regarding
  - Task complexity and nature: number of actions; structural vs. parameter changes; task duration
  - Student confusion: large branching factor=lots of trial and error steps
- Very detailed (and honest!) feedback: instructors can leverage this information

[Lins et al., SSDBM 2008]

Provenance and Teaching (2)

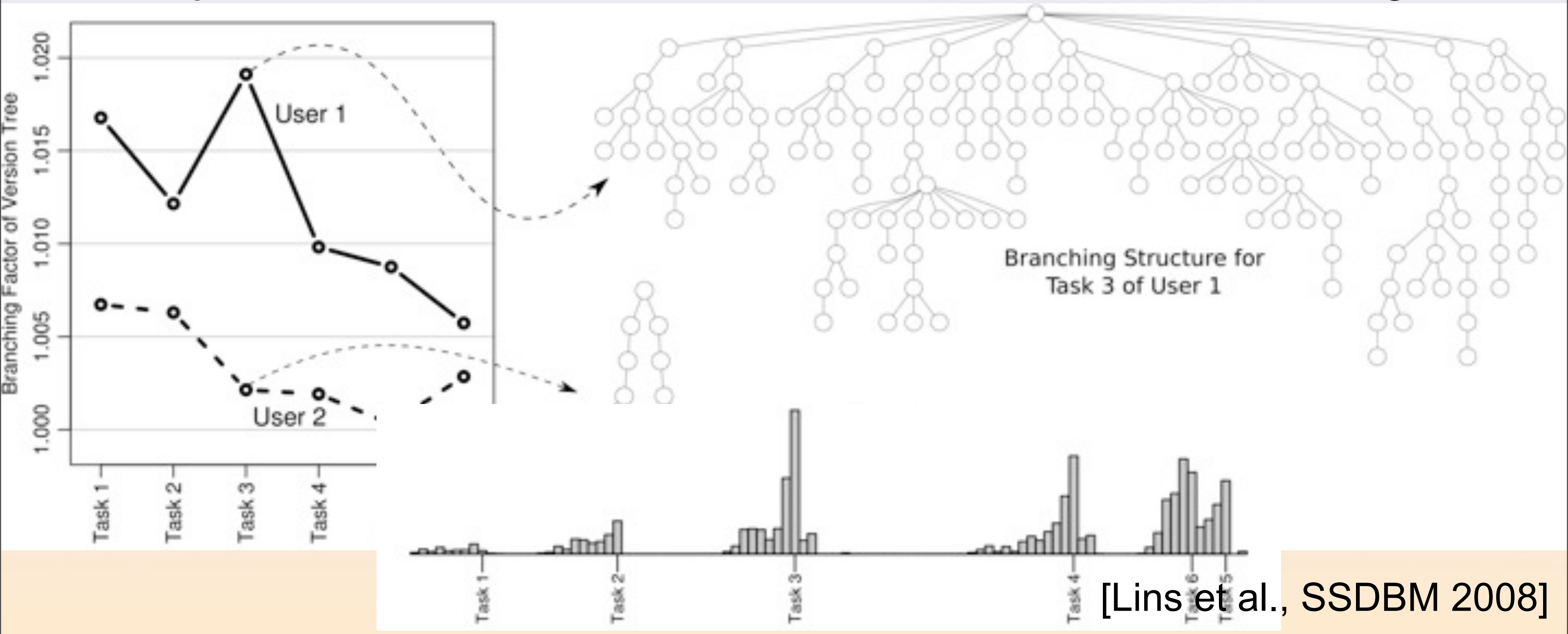
- Homework provenance provides insights regarding
  - Task complexity and nature: number of actions; structural vs. parameter changes; task duration
  - Student confusion: large branching factor=lots of trial and error steps
- Very detailed (and honest!) feedback: instructors can leverage



[Lins et al., SSDBM 2008]

Provenance and Teaching (2)

- Homework provenance provides insights regarding
  - Task complexity and nature: number of actions; structural vs. parameter changes; task duration
  - Student confusion: large branching factor=lots of trial and error steps
- Very detailed (and honest!) feedback: instructors can leverage



[Lins et al., SSDBM 2008]

Provenance and Teaching (3)

Provenance and Teaching (3)

- Homework provenance helps students and instructors to *collaborate*
  - Student is stuck, sends his provenance
  - Instructor understands student's problem, provides hints---student can see what instructor did!
  - They can also collaborate in real time [Ellkvist et al., IPAW 2008]

Using Provenance to Teach Electronic Media

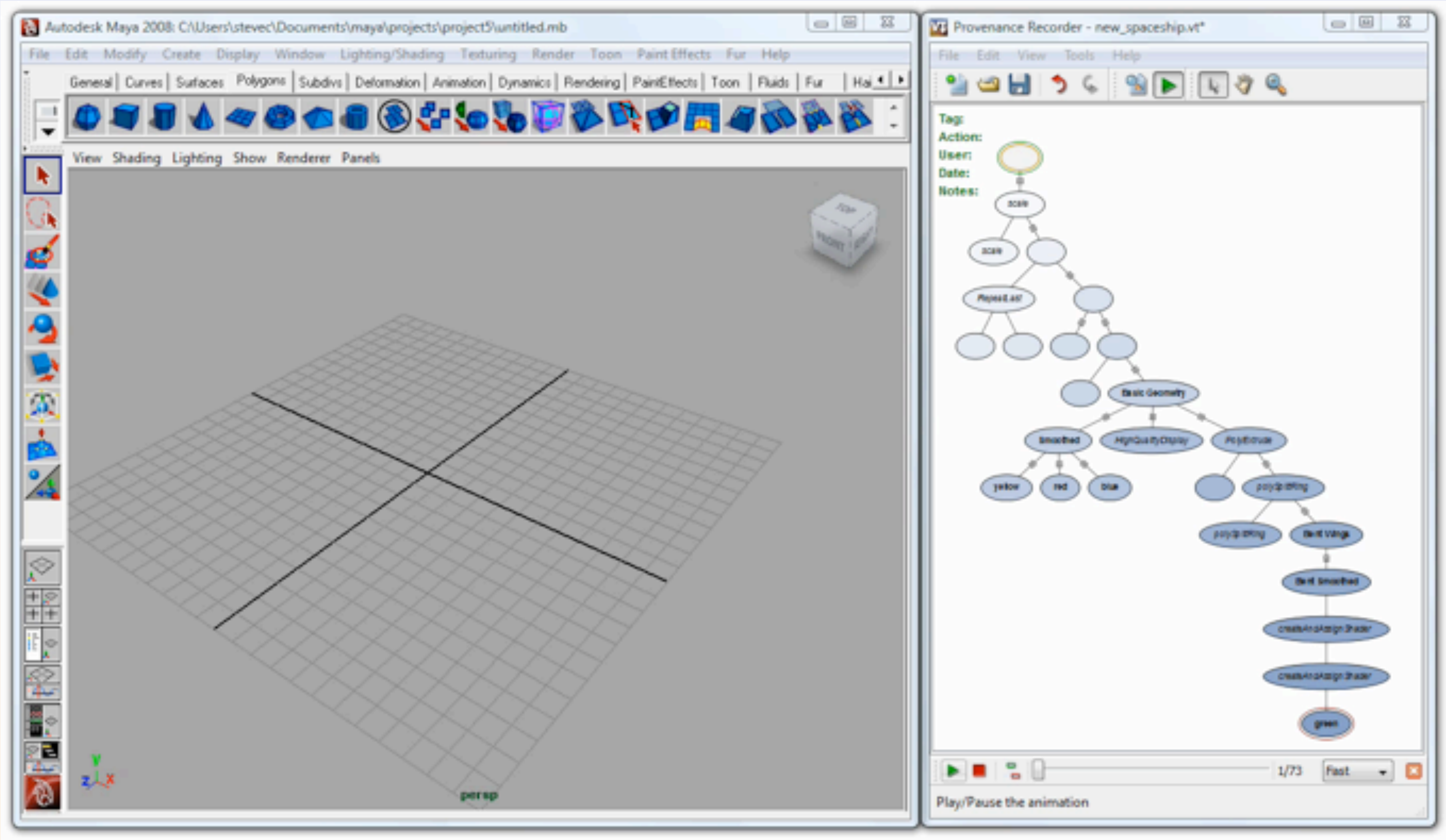
[Langefeld and Kessler, Submitted 2009]

"[...] The students have gotten to the point where they demand the VisTrails files for every demonstration just after I complete [it]"

"[...] students used [a vistrail instead of a reference model] 62% of the time"

Students who used provenance produced higher-quality models

Provenance-Based Tutorial for Maya


Autodesk Maya 2008: C:\Users\stevec\Documents\maya\projects\project5\untitled.mb
File Edit Modify Create Display Window Lighting/Shading Texturing Render Toon Paint Effects Fur Help
General Curves Surfaces Polygons Subdivs Deformation Animation Dynamics Rendering PaintEffects Toon Fluids Fur Hai
View Shading Lighting Show Renderer Panels
persp
Provenance Recorder - new_spaceship.vt*
File Edit View Tools Help
Tag:
Action:
User:
Date:
Notes:
scale
scale
Repeat as?
Basic Geometry
Smoothed
HighQualityDisplay
Polyextrude
yellow
red
blue
polySplitRing
polySplitRing
Bent Wings
Bent Smoothed
createAndAssignShader
createAndAssignShader
green
1/73
Fast
Play/Pause the animation

Provenance Analytics: Opportunities

- Volume of collected provenance is growing
- Workflow and provenance repositories
  - myExperiments (EU), Provenance Repository (Indiana), ManyEyes (IBM), Yahoo! Pipes
- Opportunity for knowledge discovery, sharing and re-use
  - Discover *workflow patterns* → a recommendation system that suggests alternatives to users as they construct a workflow
  - Discover *workflow refinement patterns* → automatically extract analogies from shared repositories
  - *Cluster* (organize) workflow collections → simplify query and search over repositories
  - *Infer workflow specification* from execution log [Aalst et al., TKDE 2004]

Provenance Analytics: Challenges

- Lots of data, complex data: graphs + metadata
  - Modules, parameters, parameter values, data products
- Do existing approaches to graph mining scale?
- Case study in clustering: [Santos, IPAW 2008]
  - Explore different workflow representations: Graphs versus bag of words
  - Examine trade-off between efficiency and cluster quality
  - Bag of words surprisingly *effective*, and *much more efficient*

[NSF Medium IIS, recommended for funding, 2009]

Conclusions and Future Work

- Advanced visualization and data analysis techniques are key to the advancement of science
- Future work into scalable algorithms, verifiable visualization, information visualization.
- Provenance management is essential for exploratory computational tasks
  - Provenance can be used to support reflective reasoning
  - Intuitive interfaces for simplifying the construction and refinement of workflows
- Science 2.0: Sharing provenance at a large scale creates new opportunities [Freire and Silva, CHI SDA, 2008]
  - Workflow/provenance repositories; provenance-enabled publications
  - Expose scientists to different techniques and tools
  - Scientists can learn by example; expedite their scientific training; and potentially reduce their time to insight
- Provenance + Workflows + Sharing have the potential to revolutionize science!

Acknowledgments

- Thanks to VGC and VisTrails group
- This work is partially supported by the National Science Foundation, the Department of Energy, an IBM Faculty Award, and a University of Utah Seed Grant.

CMOP
Center for Coastal Margin Observation & Prediction

VACET
DOE SciDAC VISUALIZATION AND ANALYTICS CENTER FOR ENABLING TECHNOLOGIES

THE UNIVERSITY OF UTAH

National Science Foundation
WHERE DISCOVERIES BEGIN

SDM CENTER

IBM

Graduate Student, Postdoc, and Software Development Positions Open